www.zte.com.cn/magazine/English

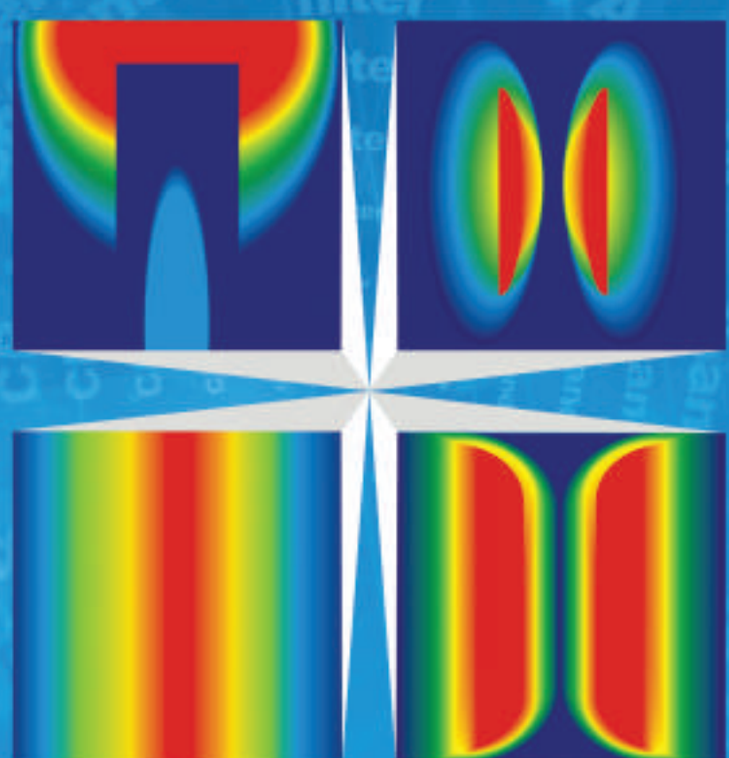
ISSN 1673–5188
CODEN ZCTOAK

ZTE COMMUNICATIONS

June 2011, Vol.9 **No.2**

Special Topic:

Microwave/RF Technologies for Future Wireless Communications



ISSN 1673-5188
06>
9 771673 518062

ZTE中兴 ZTE COMMUNICATIONS
VOLUME 9 NUMBER 2 JUNE 2011

ZTE Expects Sales Revenues from Enterprise Cloud Computing Products to Exceed $2 billion in 2011

On May 25, 2011, ZTE Corporation announced at the ZTE Analysts Conference that it hopes to venture into the IT industry by taking advantage of cloud computing. ZTE aims to pocket over $2 billion from enterprise cloud computing in 2011. Offerings will include datacom products, enterprise networks and servers, and storage products for government networks.

Gartner expects that in the global cloud computing market in 2011, the size of the PaaS + SaaS segment will be $84.63 billion, and the size of the IaaS segment will be roughly $6 billion. That means there is much space for operators to build PaaS and SaaS and to lease computing capacity and applications.

In 2011, ZTE has demonstrated a clear strategy to venture into the field of cloud computing, and to treat cloud computing as one of the new strategic growth areas for the company. Building on this, it will venture into the IT industry on a large scale. ZTE has set up a cloud computing and IT operation division and has assigned more than 3000 people to develop cloud computing solutions. The company has an entire range of Co–Cloud solutions and has already announced a series of products and solutions that include a cloud operating system, SDPaaS cloud deployment, and a cloud app platform similar to App Store. As of the end of 2010, ZTE had submitted 107 cloud–related patent applications in China. This makes ZTE the leader in cloud–related patent applications among local companies. (ZTE Corporation)

ZTE Announces Its "Uni–Evolution" Strategy

On May 25, 2011, ZTE Corporation announced its "Uni–Evolution" strategy, part of its "Network–Wide Unification" concept. The aim of the strategy is to maximize resource use through network unification, and this goal will enable operators to adapt to the marketplace with greater flexibility. The company made the announcement at its ZTE Analysts Conference 2011 in Shenzhen.

This strategy includes the entire lineup of unified solutions and covers terminals, pipes and clouds. It involves the full spectrum of access network, carrier network, application network, OAM, and terminal products. The terminal–pipe–cloud approach will become the main model for the information industry in the future.

ZTE provides unified solutions for hardware telecom equipment and telecom OAM. After successfully launching Uni–RAN and Uni–Core solutions, ZTE has also introduced Uni–NGA, Uni Bearer, and Uni–EMS for simplified and unified network structure. In terms of telecom OAM, the company has introduced Uni BSS/OSS and unified etwork management to simplify management of IT and telecom assets.

In the area of cloud computing, ZTE provides a "store on factory" solution based on "any service." ZTE has created an SaaS portal centered on a software platform similar to App Store in order to build an application platform for its operator customers. In the future, it may provide specialized enterprise–class applications for enterprise and government customers. It will also further leverage SDPaaS to realize cloud deployment and provide integrated management of IT resources in connection with its cloud operating system.

In the area of terminals, ZTE has further moved in the direction of unification. It will promote an entire range of terminals including mobile phones, PADs, data cards, IPTV, and unified app boxes for homes centered on smart terminals. (ZTE Corporation)

ZTE Wins CC Certificate from Dutch Firm

ZTE Corporation, a publicly–listed global provider of telecommunications equipment and network solutions, has become the first Chinese telecom company to be issued with a CC (Common Criteria for IT security evaluation) security certificate.

TüV Rheinland Nederland B.V., a Netherlands–based international CC security certification firm, awarded ZTE for its security, highlighting the fact that Chinese telecom companies are gaining international recognition for their IT security.

The test, which took more than six months to complete, was conducted by Brightsight, a leading IT security test and evaluation lab in the Netherlands.

It is becoming increasingly common for governments to ensure the security of telecom networks by engaging a third–party to do an evaluation.

An authorized certificate has also become one of the prerequisites for telecom equipment vendors to enter some countries and markets.

Currently, CC is the internationally recognized IT security certification. The standard adopted in this certification is IEC/ISO15408, which is also known as the CC standard, or the Common Criteria for IT Security Evaluation. CC certification is managed by the national security agency of each CC member state. There are 26 CC member states, including European countries, the United States and Japan, as well as emerging countries such as India.

CC certification includes an evaluation of security performance, loophole analysis, encryption technology and product manuals. Moreover, it also includes systematic evaluations of the entire process, from R&D and configuration management to production and shipping.

Brightsight, which performed the evaluations for ZTE, is licensed by the Netherlands, Germany and Norway and is one of the security evaluation labs recognized by countries such as India. Brightsight CEO Mr. Dirk–Jan Out said that ZTE demonstrated professionalism and strict requirements for product security in the evaluation process. He said the quick and efficient implementation of the evaluation confirmed the trustworthiness of ZTE's products.

ZTE is also applying to other CC member states to do Common Criteria security evaluations for other key products and markets. (ZTE Corporation)

ZTE Wins World's First CDMA EV–DO Rev.B Phase 2 Commercial Contract from Sistema Shyam TeleServices Ltd

ZTE Corporation announced on May 18, 2011 that it has won the world's first CDMA EV–DO Rev.B Phase 2 commercial contract from Indian telecom operator, Sistema Shyam TeleServices Ltd (SSTL). SSTL operates its telecom services across India under the brand MTS and has over 11 million wireless customers. ZTE has also won the CDMA expansion contracts and EV–DO upgrade contracts for 10 circles of SSTL. The EV–DO Rev.B Phase 2 commercial launch will commence in Rajasthan.

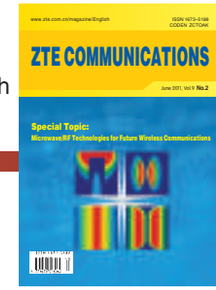
The commercial use of EV–DO Rev.B technology will greatly enhance the existing service speeds of mobile broadband for SSTL. Compared to EV–DO Rev.B Phase 1, the Phase 2 technology can further enhance BTS cell capacity, spectrum efficiency and peak rate, achieving speeds as high as 4.9 Mbit/s in single frequency carrier and 14.7 Mbit/s in a three–frequency carrier bundle. ZTE's solution based on SDR Uni–RAN technology can realize the system upgrade by just changing the channel card and software, which is able to protect SSTL's investment to the greatest possible extent.

ZTE and SSTL have been strategic partners since 2008. This win demonstrates that ZTE's CDMA technology, engineering services and other integrated capabilities have been fully recognized by SSTL.

ZTE is a global leader in CDMA technology. By March 2011, ZTE's CDMA base station capacity was 300 million lines and cumulative shipments of CDMA base stations was more than 310,000. With a CDMA market share of 32.5%, ZTE is the global leader in the CDMA market. ZTE's CDMA EV–DO Rev.B has been deployed by seven operators in China, Indonesia, Morocco, Pakistan, and Romania. In CTIA Wireless 2011, ZTE stood out with the first CDMA ultra–compact multicarrier base station, called ZTE Gecko. For this, ZTE was awarded the CTIA 2011 emerging technologies award. (ZTE Corporation)

Contents

Http://www.zte.com.cn/magazine/English
Email: magazine@zte.com.cn



ZTE COMMUNICATIONS
Vol. 9 No.2 (Issue 30)
Quarterly
First Issue Published in 2003

Supervised by:
Anhui Science and Technology Department

Sponsored by:
ZTE Corporation and Anhui Science and Technology Information Research Institute

Staff Members:
Editor–in–chief: Xie Daxiong
Deputy Editor–in–chief: Deng Xin
Executive Deputy Editor–in–chief: Huang Xinming
Editor in Charge: Zhu Li
Editors: Paul Sleswick, Yang Qinyi, Xu Ye, Lu Dan
Producer: Yu Gang
Circulation Executive: Wang Pingping
Assistant: Wang Kun

Editorial Correspondence:
Add: 450 Rongshida Avenue, Hefei 230041, P. R. China
Tel: +86–551–5533356
Fax: +86–551–5850139
Email: magazine@zte.com.cn

Published and Circulated (Home and Abroad) by:
Editorial Office of ZTE COMMUNICATIONS
Printed by:
Hefei Zhongjian Color Printing Company
Publication Date: June 25, 2011
Publication Licenses:
ISSN 1673–5188
CN 34–1294/TN
Advertising License:
皖合工商广字0058号
Annual Subscription Rate:
USD$50

Special Topic: Microwave/RF Technologies for Future Wireless Communications

Research Papers

Development Field

Operational Application

Lecture Series

Roundup

Departments

Microwave/RF Technologies for Future Wireless Communications

Ke-Li Wu and Keqiang Zhu

Ke-Li Wu is a professor of Electronic Engineering at The Chinese University of Hong Kong (CUHK). Prior to his career in CUHK he was a principal member of technical staff in the Corporate R&D division at Com Dev International. Professor Wu is a fellow of IEEE, a member of IEEE MTT-8 committee, and was an associate editor of IEEE Transactions on MTT. He has authored or coauthored more than 60 leading journal papers on EM modeling, microwave passive circuits, and antennas. His current research interests include numerical and analytical methods in electromagnetics, passive microwave circuits, microwave filters, small antennas for wireless terminals, LTCC-based multichip modules, and RF identification (RFID) technologies.

Keqiang Zhu received his B.E degree in Electronic Engineering at Hefei University of Technology in 1994. He is chief architect of RRU with ZTE Corporation. His research interests include wireless base station architecture and advanced RF technologies in wireless telecommunication industry. He has two patents in US and one patent in China.

Compared with 2G, the most prominent features of 3G and future wireless communication systems are a higher transmission rate and support for multimedia services. A higher transmission rate means that signal bandwidth is large, use of frequency spectrum is more efficient, and radio frequency equipment is greener. Demand for richer multimedia services is creating greater challenges for system developers and has led not only to the publication of tens of thousands of documents but also to tremendous new technology developments in the Long-Term Evolution (LTE) of 3GPP's Universal Mobile Telephone System (UMTS).

International deployment of UMTS is progressing steadily, and more than 180 mobile network operators throughout Europe, North America, and Asia are providing 3G services. High Speed Downlink Packet Access (HSDPA) and High Speed Uplink Packet Access (HSUPA) in UMTS can improve the transmission data rate and spectrum efficiency. This reduces transmission cost per bit. The trend towards increased data traffic and high-capacity content requires that base station equipment use new and existing frequency bands flexibly. This trend also requires that base stations have simplified but flexible network architecture with open interfaces and to consume less power. These requirements must be achieved by developing more efficient power amplifiers and broadband antennas as well as more compact high performance RF filters with less insertion loss.

In this special issue on microwave/RF technologies for future wireless communications, we invited five experts to contribute articles. Each of these articles shows a different aspect of the challenges and advanced technologies involved in microwave/RF for future wireless communications—from system architecture requirements, technologies in broadband power amplifiers, advanced RF dielectric resonator filters, and broadband antenna technologies to the state-of-the-art synthesis theory of sophisticated microwave/RF filters. We understand that these topics are far from enough to provide a complete picture of the industry, and some of the topics in this special issue are, indeed, classic. Nevertheless, we have obtained contributions from five experts, including the most experienced system architects in the industry, a senior RF engineer in power amplifiers, an industry leader who has been working on dielectric filters for more than two decades, a top tier scholar in broadband base station antennas, and the most eminent researcher in the microwave filter industry. Such combined efforts have made this issue very special.

We are very grateful to all the authors, reviewers, and the editorial board who have spent their valuable time on this special issue. We hope you will find the articles useful to your professional work and enjoyable to read.

RF Technologies and Challenges for Future MBR Systems in Cellular Base Stations

Hongyin Liao, Baiqing Zong, Jianli Wang, Keqiang Zhu, and Changjiang Cao
(Wireless Architecture Department, ZTE Corporation)

Abstract: This paper describes the advances and features of future cellular base stations. Software defined radio (SDR) evolves to cognitive radio (CR), which is smart and has wideband, and multiband radio (MBR) with reconfigurable wideband can be regarded as the basis of CR and an advanced level of SDR. Based on the SDR platform, several radio frequency (RF) solutions for implementing MBR systems are proposed, and some challenges to MBR implementation are discussed.

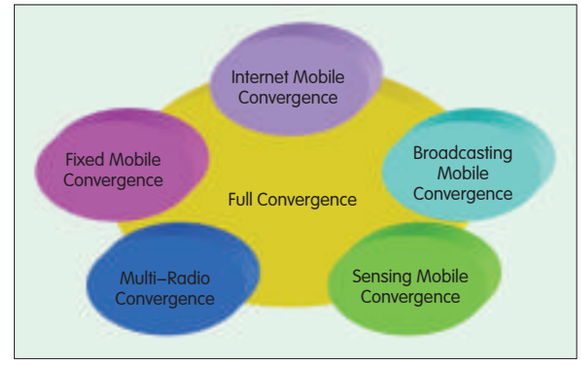
Keywords: future cellular base station; SDR; MBR; RF; challenges

1 Introduction

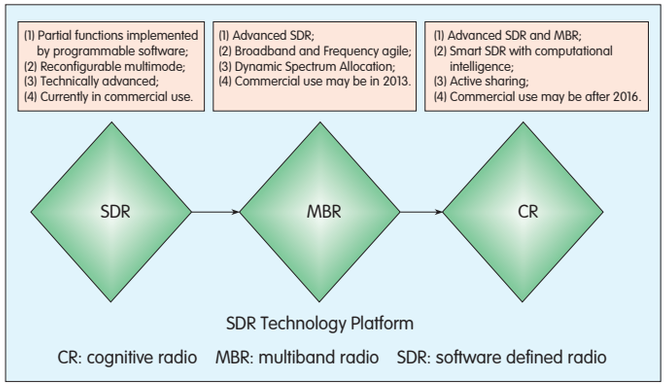
3G commercialization and evolution have been advanced and promoted by 3GPP and this has resulted in higher data rates and higher quality wireless communication services. With the development of the Internet of Things (IoT) and cloud computing, full convergence and resource sharing have become the trend of future networks (Fig. 1). In the future, mobile communication networks will be one of the primary carriers of data traffic.

With global warming and energy crises, green, energy–efficient network construction is also a key issue. This raises challenges for network operators because in a climate of competitive pressure they must optimize their investment and operating cost in order to lower CAPEX and OPEX. This is especially true for base stations. Therefore, future radio access networks (RANs) need to be smart and have advanced features such as dynamic spectrum allocation (DSA) for higher spectrum efficiency, higher performance, higher power efficiency, higher end–to–end efficiency, and all resource (active) sharing. Cognitive radio (CR) satisfies these features and has been the subject of many recent IEEE papers. At its core, a CR system can sense, adapt, and learn from its surroundings. With computational intelligence, it can also change communication parameters in response to changes in application needs or changes in the radio frequency landscape. A CR system must be wideband. Therefore, future cellular base station systems (CR systems) should be multimode, multiband, multifunctional, flexible, and smoothly upgradable. They should also be cost and power efficient, smart, and capable of coexisting and sharing equipment. Several key technologies for achieving these features are being researched [1]–[4].

Multicarrier multimode base stations are currently in commercial use and run on a software defined radio (SDR) platform. SDR is defined as a radio in which the radio frequency (RF) operating parameters of frequency range, modulation type, and/or output power can be set or altered by software, or the technique by which this is achieved [1]. An ideal SDR system allows all signals to be processed digitally except those of the analog antenna [1]. However, current SDR systems can only partly implement functions through programmable software. Bandwidth is limited by hardware, and power efficiency is not ideal.



▲Figure 1. Convergence trend of future networks.



◀Figure 2. Relationship among SDR, MBR, and CR.

SDR will evolve into multiband radio (MBR), which is a wideband system with frequency–agile devices [1]. An MBR system can change how and where devices operate within the radio spectrum, moving between a set of frequency bands in response to interference or other constraints. MBR can be called an advanced level of SDR. CR will be a more advanced technology than MBR. It will not only satisfy MBR requirements but will also be smart and have computational intelligence. CR can be seen as an expansion of SDR. Currently, radio technology is moving from SDR to MBR. As shown in Fig. 2, all SDR, MBR and CR systems run on the SDR technology platform. However, CR is smart and has computational intelligence [1]. MBR plays an important role in the development of radio technology.

2 Architecture and Features of an MBR System

An MBR system with frequency–agile characteristics can be multiband, multimode, multifunctional, flexible, and smoothly upgradable. It can also be cost and power efficient and capable of sharing equipment. Therefore, MBR can be applied to many scenarios to help operators reduce CAPEX and OPEX.

Fig. 3 shows an MBR system architecture for cellular base station. Most RF devices in this system are required to support wideband, and the duplexer or analog filter should be tunable within a frequency range of, for example, 300 MHz. It is a key point for MBR and CR to have frequency–agile characteristics. Based on the SDR platform, many MBR system functions could be performed or partly performed by a software program to reduce the difficulty in designing hardware such as RF devices.

MBR can be classified into two levels according to the range of frequency agility. One level is reconfigurable within a frequency band range of 300 MHz; that is, 700–1000 MHz. The other (higher) level is reconfigurable within a frequency band range of no less than 1 GHz; that is, 1–2 GHz. The system design for the latter is more difficult than for the former. On every level, subsystems except the duplexer allow wideband from 0 Hz to the maximum bandwidth of, for example, 300 MHz or 1 GHz. The duplexer comprises multiband filters, and every sub–band filter satisfies the protocol requirements of the communication mode in that sub–band.

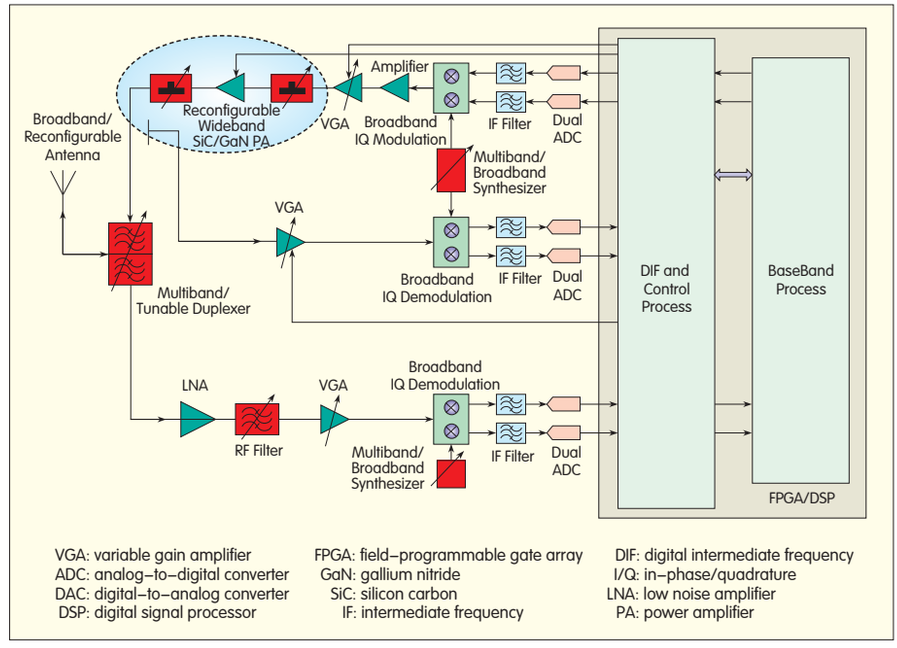
MBR is a whole solution that caters for the advancement of network. However, in implementing key MBR technologies such as multiband/broadband antenna, multiband duplexer, broadband/multiband filter, high efficiency broadband power amplifier (PA), broadband synthesizer, high speed ADC/DAC converters, and high performance field–programmable gate array (FPGA) or digital signal processor (DSP), there are challenges. The cost of implementing MBR should also be carefully considered.

3 Key RF Technologies and Challenges for an MBR System

An MBR system is wideband and based on the open SDR platform. Therefore, the following discussion also applies to SDR systems.

3.1 Broadband/Reconfigurable Antenna

The antenna in an MBR system must be broadband and multiband for port sharing, providing broadband



▲Figure 3. Architecture of an MBR system.

coverage, and providing flexibility. Most currently used multiband antennas are multiband and multiport (or a port with a combiner inside). Although the antenna provides dual/triple/multiband performance, every sub–band corresponds to a port.

The broadband and multiband antenna design of cellular base station antennas is based on microstrip antenna technology. Microstrip technology has a planar electric dipole and a shorted patch antenna (equivalent to a magnetic dipole) for achieving a wide range of voltage standing wave ratio (VSWR) performances and operating bandwidth with excellent electrical characteristics [2], [3]. Broadband load matching with excellent electrical characteristics is a challenge to antenna design. Moreover, cost, size, and weight of antennas are also of concern to operators.

Broadband and multiband antennas covering about 300 MHz have been released by vendors such as Andrew. However, implementing such antennas in outdoor base stations covering about 1GHz and achieving excellent electrical performance is a challenge with present technology. Mobile Mark Inc. has released an antenna called the Surface Mounted Multiband (SMW) antenna. It is a small distributed system antenna that has many different indoor wireless applications. It has two broadband antenna elements: an 800–2700 GHz element that can be used for Cellular 850/1900/2100 MHz, WiMAX 2.5 GHz, or a second Wi–Fi, and a 1700–2700 MHz element that can be used for Advanced Wireless Services (AWS–1) band 1.7–2.1 GHz, GSM 1.9 GHz, Wi–Fi 2.4 GHz, and WiMAX 2.5 GHz.

One approach to this challenge is to use reconfigurable antenna technology based on RF Micro–Electro–Mechanical System (MEMS) [5]–[9]. This technology is intelligent and state of the art. A reconfigurable antenna design using a network of MEMS switches can change its operating frequency and radiation/polarization characteristics, which are the goals of MBR. Barriers to implementing RF MEMS reconfigurable antennas in cellular base stations are performance, reliability, power limit of RF MEMS switches, design of switch bias networks, and control algorithms [8], [9].

3.2 Multiband/Tunable Duplexer

In a current SDR system, the duplexer works in a single band. A 900 MHz frequency band duplexer for GSM\UMTS\LTE has RX of 890–915 MHz, TX of 935–960 MHz, filter bandwidth of 25 MHz, and total bandwidth of 70 MHz. The maximum bandwidth of the duplexer for GSM\UMTS\LTE 1800 MHz is about 170 MHz (with 75 MHz filter bandwidth). However, a multiband duplexer should be tunable to no less than 300 MHz with sub–band filter covering no less than 40 MHz for an SDR base station. Therefore, the multiband duplexer may be the most difficult component to design in an MBR system.

For wireless duplexer applications, TX and RX filters are required to have extremely sharp roll–offs. Current base stations mostly use high–Q air cavity duplexers in which multiple cavities are in series with TX OUT or RX OUT for high performance. A duplexer cavity comprises resonant cavities which can simply be two carefully tuned resonant circuits. One circuit sets the bandpass frequency the cavity is resonate on, and the other is for coupling energy into the cavity. Respectively, these circuits have a bolt for tuning the bandpass frequency and for tuning the frequency of the cavity notch.

The difficulty in designing a reconfigurable multiband duplexer for an MBR system lies in mechanical requirements and achieving excellent electrical characteristics with frequency shift and variable sub–bandwidths. Currently solutions only use a motor to tune the bolt depth or move coppers in the top covers of all cavities for resonating on the desired bandpass frequencies. Also, they only cover 200 MHz bandwidth for narrower sub–band application (Fig. 4). High cost is also a key obstacle for applying multiband duplexers.

The novel duplexer architecture being researched in [10] may be an ideal solution for reconfigurable multiband duplexers in MBR system. This solution can be used to replace the duplexer or reduce its strict performance so that it can be tunable easily. The architecture combines a low isolation device with an adaptive loop canceling scheme and is same as feedforward fashion. It provides the required transmitter leakage and transmitter noise isolation over wideband by using a delay element and an adjustable vector attenuator in cancellation path (Fig. 5) [10]. The feasibility of wideband cancellation depends mainly on the delays in the main path and cancellation paths. These delays are restricted by attenuation coefficients in the cancellation paths. The achievable cancellation bandwidth also depends on the duplex frequency; smaller delay differences and smaller duplex

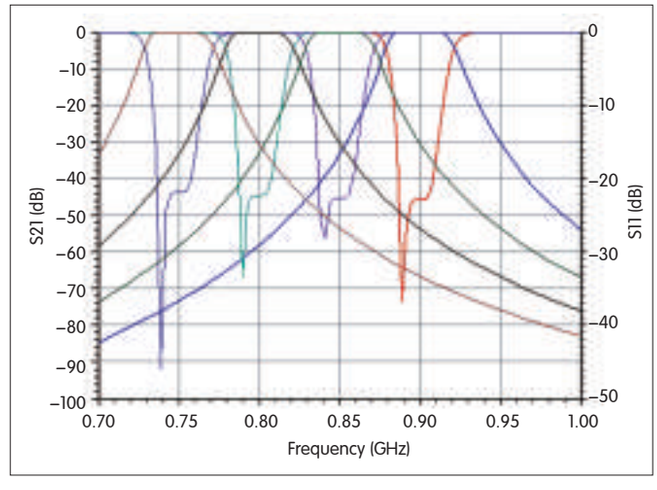


Figure 4. Performance of a reconfigurable duplexer from simulation.

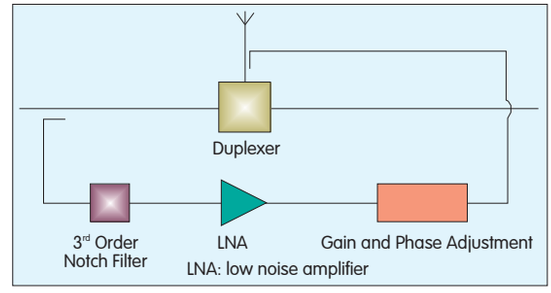
frequencies give higher cancellation bandwidths. Wider bandwidth cancellation can also be achieved by employing two or more loops. Arithmetic and noise floor from RF devices may be a more important factor that others affecting cancellation. The test results are not very ideal and further study is needed [10].

A more advanced technology for application in multiband duplexers is metamaterial duplexer technology. This is currently in the theoretical and lab testing stage. When applying this technology in multiband duplexers, the left–hand artificial structure metamaterial with negative refractive index exhibits unusual properties. If RF signals input it with different injection angle, different pass band and stop band which is the want of designing multiband duplexer will appear [11].

3.3 Wideband PA Technologies

For high efficiency and broadband connectivity in an MBR system, the PA in the system should cover wide bandwidth and have a high linearity for amplifying signals without distortion. These performances require transistors with higher impedance and allow for easier and lower loss matching networks in amplifiers. Purely real impedances can theoretically be matched to a 50 Ω system over any bandwidth by using an infinite number of matching elements. However, actual devices have optimum impedances with a reactive component. Complex loads can be matched only over a limited bandwidth as defined by Fano's limit [4]. A suitable figure of merit for high power broadband capability in a device technology is a low pF/W gate and drain capacitance. Therefore, a wide bandgap (WBG) semiconductor such as gallium nitride (GaN)—which can be operated at high drain voltage and has low parasitic capacitances per watt of output power—is a favorable choice for use in frequency–agile pulsed applications such as military radar, air traffic control radar, and communications jamming [4].

In current SDR systems, laterally diffused metal oxide semiconductor (LDMOS) is the most important technology for the PA. Its large device parasitic capacitances per watt of output power lead to low device input/output impedances. However, because of the narrow instantaneous bandwidth (about 40 MHz) and gain characteristic of LDMOS devices, it is no longer fit for MBR systems. GaN technology has been developing fast and will gradually be commercialized. The bandwidth of GaN RF PA transistors is generally more than 300 MHz, even GHz in microwave frequency band. There are several circuit architectures for GaN PA applications, including Class AB, Doherty, and very–high–efficiency Class D/E/F/F–1. The leading vendors of GaN PA transistors include RFMD and CREE. Performance, cost, and reliability are the main challenges to GaN PA application [4].



▲Figure 5. Block diagram of a novel duplexer architecture.

Digital predistortion (DPD) plays an important role in an SDR system; it improves PA efficiency and reduces the difficulty of designing hardware devices. A accurate non–linear RF PA transistor model ensures high performance of DPD, and LDMOS device models are rich and perfect. However, GaN RF is an emerging technology, and research on non–linear modeling is still challenging.

Wideband Transceiver Technologies

In an MBR system, a transceiver should support wideband and should have reconfigurable frequency band. There are conflicting requirements on the transceiver, including wideband, multimode, high dynamic range, high power efficiency, cost and size. In current SDR systems for cellular base stations, the transceiver architecture is mainly based on a direct or single conversion solution with wideband performance at zero or high intermediate frequency. As shown in Fig. 3, the challenges for this MBR system architecture lie mainly in the design of the ADC\DAC, multiband synthesizer, and image filter. High dynamic range and high resolutions of ADCs\DACs (such as 14–bit ADC and 16–bit DAC) can be used in the direct or single conversion transceiver architecture in order to satisfy the systemic requirements. Several multiband synthesizers have been released, including the AD4350 (which covers 137.5–4400 MHz), and the HMC22 (which divides RF output into three bands: 665–825 MHz, 650–1330 MHz, and 2660–3300 MHz). The frequency band in the image filter is variable within the required range. Combining multiband filters with switches is a current solution for the image filter in multimode and multiband handsets. However, this solution is large and expensive and is not flexible and advanced.

With low power consumption, high isolation, high density, and high integration advantages, RF MEMS is an emerging area in tunable filter design. By changing the values of MEMS switches or varactors within RF filters, tunable characteristics can be achieved [5]. Film bulk acoustic wave resonator (FBAR) filter techniques are also a current area of research interest. FBAR devices have lower loss (high Q value), better power–handling, and better robustness with the most demanding specifications. Because they are more expensive to manufacture than other solutions, there is still much research to be done [12].

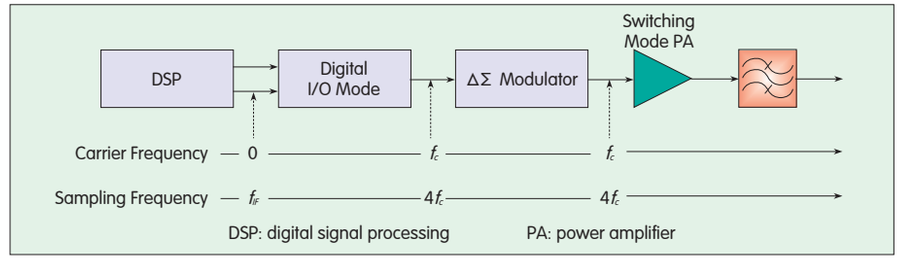
In Fig. 3, the transmitter and feedback receiver have no filters because the DPD bandwidth is too wide, and the algorithms such as DPD and Auto Quadrature Error Correction (AQEC) in baseband or digital intermediate frequency help reduce spurious signals. Accordingly, the difficulty in designing analog devices is reduced.

The architecture for a zero

intermediate frequency (ZIF) receiver has a pathway for full on–chip integration of the receiver because the signal is directly demodulated to baseband I and Q signals. Because the intermediate frequency (IF) is zero, there is no need for an external IF surface acoustic wave (SAW) filter for channel selection, and there is no need for an additional IF synthesizer section or an image reject RF filter (required in a high IF receiver). The basic RF filter in the duplexer rejects out–of–band blockers and transmitter leakage (Fig.3). However, an external bandpass RF filter might still be required after a low noise amplifier (LNA) in order to further reject out–of–band blockers and transmitter leakage at the demodulator input caused by limited finite duplexer TX–RX isolation. Channels are selected at baseband by on–chip low–pass filters that filter channels and reject close–in blockers. The bandwidth of on–chip baseband filters can be programmed on–chip so that the filters can operate the receiver in multiband and multimode applications. After channel filtering in a ZIF receiver, I/Q signals at baseband are amplified by variable gain amplifiers before they are digitized in the analog baseband section. Challenges to designing a ZIF receiver include high IIP2 with the second–order distortion, and the algorithm on current AQEC. The available IF bandwidth of this solution is about 20 MHz and cannot satisfy the requirements of multicarrier GSM because of the dynamic ADC range [13].

Using flexible radio architecture on an SDR platform in order to support an MBR system and multiple wireless standards has attracted much interest. The digital design flow enables a higher level of system integration and higher bandwidth, simplifies testability, and provides reduced power, size, and cost. Some advanced transceiver architectures that have not been put into commercial use will be introduced here. These include RF sampling receiver and all digital transmitter [13]–[23].

The digital transmitter with switching mode PA (SMPA) is an attractive choice for current SDR systems and next generation MBR systems because of its high power efficiency, linearity, low complexity, flexibility, reconfigurablity, and wideband. Much research has been focused on all–digital transmitter architectures such as low–pass delta–sigma modulator (DSM), bandpass DSM, and direct–pulsewidth/position–modulation (PWPM) [13]–[18]. Theoretically, all–digital transmitter uses a quantizer at the modulator's output to generate a pulse–shaped signal. The quantization noise is spread over a wide band and is shaped outside the useful band of the signal using interpolation and the delta–sigma transfer function. This architecture generates an entirely digital two–level signal at RF, so it is configurable and suitable for multistandard and multiband applications. Fig. 6 shows novel bandpass DSM transmitter architecture [18].



▲Figure 6. Novel bandpass DSM transmitter architecture.

The DSM transmitter for wireless RF applications has the following two drawbacks:

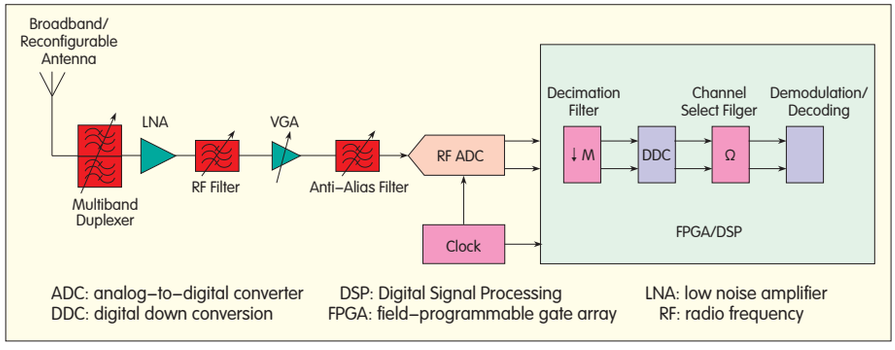
(1) RF signals are centered at several gigahertz, and oversampling of the carrier requires enormous digital clock rates. This limits the signal bandwidth to 25 MHz based on currently available technologies [13]–[18].

(2) The architecture is complex. The modulator needs to work at several gigahertz and requires high–speed computational capability for digital signal processor or FPGA. This greatly increases the cost and power consumption of the designed circuits. These challenges give direction to future research.

It has recently been popular to use RF–sampling techniques for implementing more flexible front–ends in terms of frequency tuneability, filtering, and easing of requirements on the ADC [19]–[23]. A receiver with this RF–sampling ADC is most compact and has a very simple design, no LO and mixer, and a relaxed filter. But its drawbacks are evident—it requires more digital processing, is has the highest sampling rate, it is most sensitive to jitter, require a higher power converter, and its overall linearity is limited by ADC. Fig. 7 illustrates a direct RF sampling receiver architecture with wide frequency or multiband coverage and arbitrary tuning. This architecture provides a high degree of reconfigurability of tuning range and bandwidth by using a tunable or selectable anti–aliasing filter before the stage of RF conversion [19]–[23]. The sampling rate of ADC for direct RF bandpass conversion receiver is defined by a tunable bandwidth of, for example, 300 MHz. This is at least double the bandwidth for meeting the Nyquist law, so it is not extremely high. The difficulty with this type of ADC application is that, currently, its dynamic range cannot meet the requirements of cellular base stations. Companies such as ADI and TI have been researching and developing this technology. It is conceivable that this RF–sampling technique will be in commercial deployment soon. Another type of ADC architecture is low pass sampling and oversampling. This has a very high sampling clock of about several gigahertz if applied in RF band, for example, level two MBR scenes. This architecture is now in a technologically ideal state.

4 Conclusions and Prospects

MBR with frequency–agile



▲Figure 7. An architecture of direct RF sampling receiver.

characteristics can be considered the advanced stage of SDR and the basis of CR. MBR is a wideband and frequency–agile system that is implemented on the SDR platform. Based on MBR for cellular base stations, GSM, UMTS, LTE, and other communication systems can be smoothly upgraded, can coexist, and can share a number of network elements. However, there are still several technical challenges to MBR implementation and commercial deployment. These will require further study.

Because of limited hardware bandwidth, power inefficiency, size and cost of radio systems, as well as improvements in DSP and FPGA, analog processing is being turned into digital processing using digital compensation and algorithms. The compensation solutions presented in this paper allow for an easing of analog requirements for small, low cost, flexible and highly reconfigurable radios in broadband communication systems. This trend is inevitable and attractive.

Acknowledgement:

Thanks to Wei He from the Shanghai RRU department of ZTE Corporation for his contribution to this work.

References

[1] T. Weingart, "A method for dynamic reconfiguration of a cognitive radio system," Ph.D dissertation, Dept. Comp. Sci., Uni. Colorado, Boulder, 2006.
[2] T.G. Spence and D.H. Werner, "A novel miniature broadband/multiband antenna based on an end–loaded planar open–sleeve dipole," *IEEE Trans. Antennas Propag.*, vol.54, no12, pp.3614–20, 2006.
[3] Zhi Ning Chen and Kwai–Man Luk, "Antennas for Base Station in Wireless Communications", New York: McGraw Hill, 2009.
[4] S. Azam, C. Svensson and Q. Wahab, "Comparison of two GaN transistor technologies in broadband power amplifiers", *Microwave Journal*, 53(4), p.184, 2010.
[5] R. J. Richards and H. J. De Los Santos "MEMS for RF/Microwave Wireless Applications: The Next Wave," *Microwave Journal*, March 2001.
[6] B. Cetiner, H. Jafarkhani et al, "Multifunctional reconfigurable MEMS integrated antennas for adaptive MIMO systems," *IEEE Commun Mag*, vol. 42, no. 12, pp. 62–70, 2004.
[7] T. Ativanichayaphong, Ying Cai, Jianqun Wang, Mu Chiao and J.Chiao, "Design considerations of reconfigurable antennas using MEMS switches," in *Proc. SPIE Microelectronics, MEMS, and Nanotechnology Symp.*, Brisbane, Australia, 2005.
[8] J. Bernhard, "Reconfigurable antennas and apertures: state of the art and future outlook," in *Proc. SPIE Conf. Smart Electronics, MEMS, BioMEMS and Nanotechnology*, San Diego, CA, 2003, pp.1–9.
[9] B. Cetiner, H. Jafarkhani, J. Qian et. al, "Multifunctional reconfigurable MEMS integrated antennas for adaptive MIMO systems," *IEEE Commun. Mag.*, vol.42, pp.62–70, Dec. 2004.
[10] T. O´Sullivan, R. York, B. Noren, P. Asbeck, "Adaptive duplexer implemented using feedforward technique with a BST phase shifter," *IEEE Trans. Microw.Theory Techniques*, vol. 53, no 1, pp. 106–114, Jan. 2005.
[11] S.Geelani, "High Q–factor metamaterial duplex filters in suspended stripline technology," Ph.D dissertation, Technischen Fakultät, Universität Erlangen–Nürnberg zur Erlangung des Grades, Erlangen, März, Germany, 2009.
[12] C. H. Tai, T. K. Shing, Y. D. Lee and C. C. Tien, "A novel thin film bulk acoustic resonator (FBAR) duplexer for wireless applications," *Tamkang J. Sci. Eng.*, vol. 7, no. 2, pp. 67–71, 2004.
[13] W.Y Ali–Ahmad, "Radio transceiver architectures and design issues for wideband cellular systems," *Proc. IEEE Int. Workshop Radio Freq. Integration Tech.*, Singapore, p.21, 2005.
[14] M.Helaoui, S. Hatami, R. Negra and F. Ghannouchi, "A novel architecture of delta–sigma modulator enabling all–digital multiband multistandard RF transmitters design," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol.55, no.11, pp.1129–1133, Nov. 2008.
[15] V. Parikh, P. Balsara and O. Eliezer, "All digital–quadrature–modulator based wideband wireless transmitters," *IEEE Trans.Circuits Syst. I, Reg. Papers*, vol.56, no.11, pp.2487–2497, Nov. 2009.
[16] Sung–Rok Yoon and Sin–Chong Park, "All–digital transmitter architecture based on bandpass delta–sigma modulator," *Int. Symp.Commun.Inf. Tech (ISCIT)*, Incheon, Korea, pp.703–706, 2009.
[17] B. Thiel, S. Dietrich, N. Zimmerman and R. Negra, "System architecture of an all–digital GHz transmitter using pulse–width/position–modulation for switching–mode Pas," *Asia Pacific Microwave Conf. (APMC)*, Singapore, pp. 2340–2343,2009.
[18] A. Jayaraman, P. F. Chen, G. Hanington, L. Larson, and P. Asbeck, "Linear high efficiency microwave power amplifiers using bandpass delta–sigma modulators," *IEEE Microwave and Guided Wave Letters*, vol. 8, pp. 121–123, Mar. 1998.
[19] Johann–Friedric Luy, T Mueller, T. Mack, A Terzis, "Configurable RF receiver architectures," *IEEE Microwave Magazine*, pp. 75–82, March 2004.
[20] D. M. Akos, M. Stockmaster, J.B.Y. Tsui, J. Caschera, "Direct bandpass sampling of multiple distinct RF signals," *IEEE Transactions on Communications*, pp. 983–988, July 1999.
[21] Gerald L. Fudge, "Reconfigurable direct RF bandpass sampling receiver and related methods," United States Patent Application # 20070081617, 12 April 2007.
[22] R. Barrak, A. Ghazel, F. Ghannouchi, "Design of sampling–based downconversion stage for multistandard RF subsampling receiver," IEEE Int. Conf. on Electronics, Circuits and Systems, pp. 577–580, 10–13 Dec. 2006.
[23] M. L. Psiaki, S. P. Powell, H. Jung and P. M. Kintner, "Design and Practical Implementation of Multi–frequency RF Front Ends Using Direct RF Sampling," IEEE Trans. on Microwave Theory and Techniques, vol. 53, Issue. 10, pp. 3082–3089, Oct. 2005.

Biographies

Hongyin Liao (liao.hongyin@zte.com.cn) received his M.E. and Ph.D. degrees in radio communication from Shanghai University. He currently works for ZTE Corporation as a system architect for wireless base stations. His research interests include wireless communication and digital/RF technology. He has published more than 10 papers and holds a patent in China.

Baiqing Zong (zong.baiqing@zte.com.cn) received his B.E. and M.E. degrees at Nanjing University and Peking University. He received his Ph. D. degree from Zhejiang University in 1998. He currently woks for ZTE Corporation as an architecture engineer of wireless products. His research interests include mobile communication and RF technologies. He has published more than 10 papers.

Jianli Wang (wangjianli@zte.com.cn) received his M.E. and Ph.D. degrees in precision instruments at Tianjin University. He is currently a chief engineer at ZTE. His research interests include wireless system architecture and RF technology.

Keqiang Zhu (kzhu@zteusa.com) received his B.E degree in electronic engineering at Hefei University of Technology in 1994. He is chief architect of RRU with ZTE. His research interests include wireless base station architecture and advanced RF technologies in wireless telecommunications. He has two patents in the U.S. and one patent in China.

Changjiang Cao (cao.changjiang@zte.com.cn) received his M.E. degrees in electronic mechanisms at Xidian University. He received his Ph.D. degree in automatic control from Shanghai Jiaotong University in 2001. He is currently a senior wireless architecture engineer at ZTE Corporation. His research interests include wireless network structure and active antenna technology. He has published more than 10 papers.

Broadband Power Amplifiers for Unified Base Stations

Pengcheng Jia
(ZTE USA)

Abstract: A broadband power amplifier is required to cover the full range of cellular frequency band—from 700 MHz to 2600 MHz—in a base station that supports multiple frequency bands simultaneously. Conventional laterally diffused metal oxide semiconductor (LDMOS) transistors support narrow band applications up to 3 GHz. However, they cannot operate beyond 1 GHz in broadband applications. GaN transistors have much higher power density and operational frequency compared with LDMOS. Therefore, they are ideal for broadband amplifiers that support multiple bands. Theories for designing broadband amplifiers are introduced in this article, and a 500–2500 MHz 60 W GaN amplifier is discussed.

Keywords: broadband power amplifier; GaN; LDMOS

1 Introduction

It is very challenging for multinational carriers to support multistandard multiband networks. A base station has to support frequencies such as 800 MHz, 900 MHz, 1800 MHz, 1900 MHz, 2100 MHz, and 2600 MHz. It is costly to install independent Radio Frequency (RF) modules that support each frequency. Therefore, a broadband RF module has attracted the attention of carriers because it not only reduces the size of the RF module but is also more energy efficient. It can also lower operation and maintenance cost.

In an RF module, the most challenging component is the power amplifier. It is very difficult to develop a broadband amplifier that can work beyond 1 GHz with conventional laterally diffused metal oxide semiconductor (LDMOS) transistors. Emerging GaN technology is perfectly suited for this broadband application. A GaN transistor has much higher power density and operational frequency compared with LDMOS, which makes it a good candidate for broadband power amplifiers.

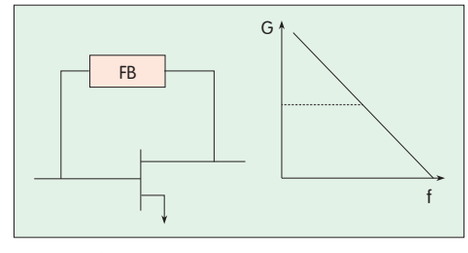
2 Designs of Broadband Power Amplifiers

2.1 Feedback Amplifier

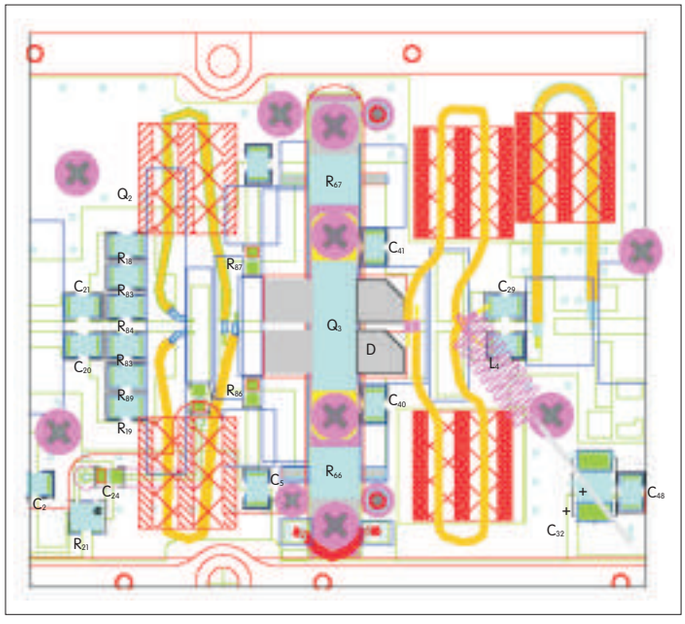
LDMOS transistors are currently the dominant power transistor candidate for base station power amplifiers. Their output power is high, and their cost is very attractive. However, the peripheral of LDMOS transistors is normally large, and its optimal load is typically in the sub–Ohm range. The high impedance transformation ratio (from sub–Ohm to 50 Ω) limits the bandwidth of the matching network. Typically, it can only cover 100 MHz of bandwidth or less.

Feedback is a classic approach to broadening the bandwidth of an amplifier [1]. As shown in Fig. 1, a feedback can be placed between the drain and gate of a LDMOS transistor. Generally, the gain of the transistor increases monotonically when the frequency drops. With a feedback, the gain curve may be flattened below the corner frequency. The input and output match can both be improved by the feedback when its value is properly selected. The intermodulation components can also be suppressed by the feedback.

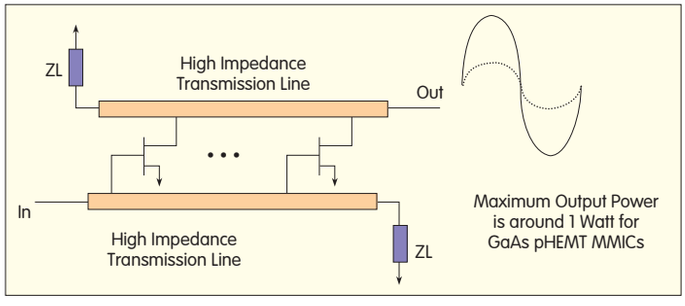
A broadband LDMOS power amplifier with a high power feedback structure is shown in Fig. 2. Two transistors are assembled inside a ceramic package. Feedback resistors are placed on the top and bottom of the package. The feedback resistors are connected to each transistor by lines around the transistor. Capacitors are placed in the feedback path to isolate the DC voltage from drain to gate. The feedback resistor has a metal flange that is directly mounted to the base plate. The feedback resistor can dissipate a large amount of heat, and the length of the feedback path is



▲Figure 1. Feedback amplifier and its gain curve.



◀Figure 2.
LDMOS feedback amplifier.



◀Figure 3.
A distributed amplifier.

minimized to avoid oscillation.

From the output port, a 2:1 balun is first used to split the 50 Ω port impedance into two 25 Ω channels. Then, a 4:1 balun is used to further transform the impedance from 25 Ω to 6.25 Ω. It is much easier to match the optimal load of the transistor to 6.25 Ω. A similar balun is used at the input to improve the matching at the input side. This feedback amplifier can output more than 60 W of power from 20 MHz to 1000 MHz.

2.2 Distributed Amplifier

Another common approach to design a broadband power amplifier is the distributed amplifier, as shown in Fig. 3. It comprises a series of small transistors. The gates of field–effect transistors (FETs) connect to the input transmission line, and their drains all connect to the output transmission line. The distance between transistors is properly designed so that the output power from each transistor is summed in phase. The characteristic impedance of the input/output transmission lines is higher than 50 Ω. When it is loaded periodically with a series of transistors, its characteristic impedance is lowered to 50 Ω and presents a good input/output match.

A distributed amplifier has a very broad bandwidth and can be applied to millimeter wave applications. However, its output power is limited by the breakdown voltage of the last stage transistor, and its efficiency is not good either. Half of the energy is dissipated into the dummy load ZL on the output transmission line. The transistors are typically biased at class A state, and the overall efficiency of the distributed amplifier is only 25% theoretically. It is not a good candidate for base station application because of its poor efficiency and low output power.

2.3 In–Phase Power Combining

Higher power can be achieved by combining the output power from more than one amplifier cell by broadband combiners [2]. A multisection broadband Wilkinson combiner is shown in Fig. 4. The combiner has four sections and covers a 4:1 (fmax:fmin) frequency range.

The combiner network has two branches, and each branch transforms impedance from the summing point (100 Ω for a two–way combiner) to the amplifier port impedance (typically 50 Ω). Each branch has several sections of transmission lines with each section around quarter wavelength at the center frequency. The characteristic impedance of each section changes gradually, and the section closest to the amplifier has the lowest impedance. In combiner design, the network's Q factor can be defined by the impedance transformation ratio Qr for each transmission line section. The lower the Qr , the broader the

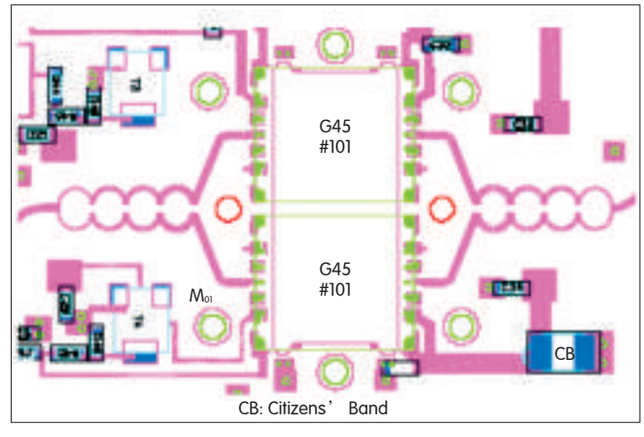
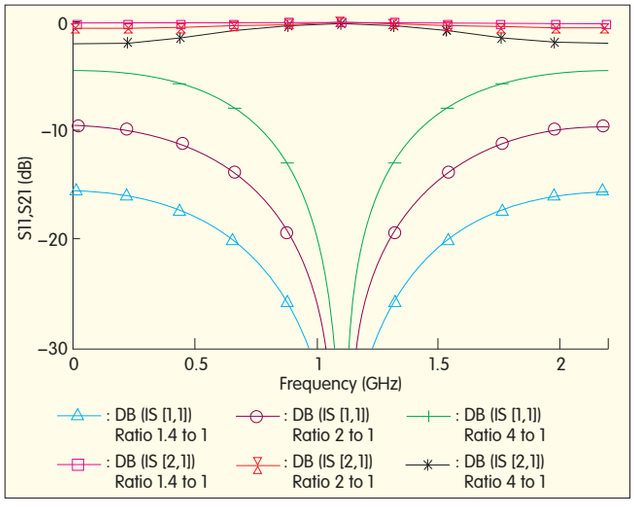


Figure 4. ▶
Multi–section Wilkinson combiner.



◀Figure 5. Bandwidth comparison for different impedance transformation ratios.

bandwidth. To broaden the bandwidth, more sections of transmission lines should be used to reduce the impedance transformation ratio at each section. The relationship between Qr factor and bandwidth is shown in Fig. 5.

2.4 Power Combining with 90 Degree Hybrids

More output power can also be achieved by combining two amplifier cells with a pair of 90 degree hybrids (Fig. 6). Each of the amplifier cells has the same input/output matching network. The hybrid power combining approach offers the benefit of wide bandwidth. The amplifier is unconditionally stable within the pass−band of the hybrid. It also provides good input/output match. The reflected signals from the inputs of both amplifier cells sum at the input port of the hybrid. One reflected signal goes through the 0 degree coupling port twice, and the other reflected signal goes through the 90 degree coupling port twice. The two reflected signals are out of phase and cancel each other at the input port. This gives the amplifier a very low input and output return loss. When designing a multistage amplifier, this helps reduce the gain ripple, and the driver amplifier's output power can be fully delivered to the output stage.

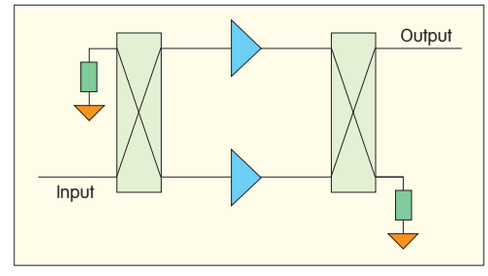
3 GaN Broadband Power Amplifier Demonstration

Typically, high power transistors have very low optimum source and load impedance at very high output power level. The GaN transistor can be operated at much higher voltages; for example, 48 V, compared with 28 V for LDMOS. The load impedance is much higher for a GaN transistor when operating at higher voltages compared with the sub−Ohm load impedance for conventional transistors such as GaAs FET or Silicon LDMOS. The peripheral of a GaN transistor is also much smaller than that of a LDMOS transistor with the same output power capability. This leads to a much smaller parasitic capacitance and a simpler matching network design. The GaN transistor also has very low thermal resistance when it is grown on SiC substrate. Its operating frequency can increase to millimeter wave range because of much higher electron mobility. These features make the GaN transistor an ideal candidate for broadband applications.

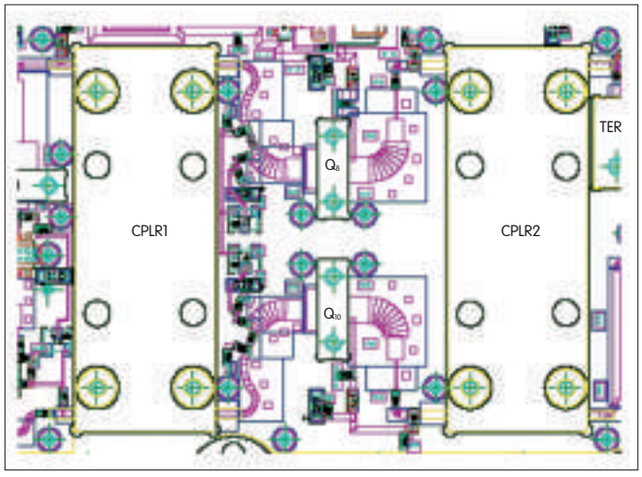
GaN transistors have much higher gain at low frequency and are very easy to oscillate. A feedback network can help control the gain at lower frequency. However, because of the physical length of the feedback network, it can change into positive feedback at higher frequency. So the feedback network must be carefully selected. It needs to be inductive and become a high impedance component at the high frequency end. Heat dissipation also needs to be taken intr high power amplifier designs. The feedback resistor consumes power and needs to be capable of dissipating heat. With a proper feedback network, a GaN amplifier can operate over an extraordinarily broad bandwidth, ranging from 20 MHz to 2500 MHz with an output power of more than 20 W.

For very high power applications, amplifiers are typically designed without a feedback network. The bias network and matching network has to be carefully selected to compress the gain at low end and avoid oscillation. Fig.7 shows a GaN amplifier with a broad bandwidth covering 0.5 to 2.5 GHz frequency. The output stage uses two GaN transistors, which can output 90 W at narrow band applications at 32 V. The transistor is currently offered in die format only. A flange−type package is selected, and the die is mounted into the package first. Then, the packaged GaN transistor is dropped into the power amplifier module. The output stage uses a pair of broadband 90 degree hybrids to combine the output power of two transistors. The hybrid is a strip−line−type 90 degree coupler. It has multiple sections of strip line networks stacked between a top and bottom aluminum plate. The hybrid is assembled into the amplifier module through cavities in the printed circuit board (PCB), and the leads of the hybrid are soldered to the microstrip line pads. The matching networks for each transistor are identical. A good input/output match can be achieved for the output stage.

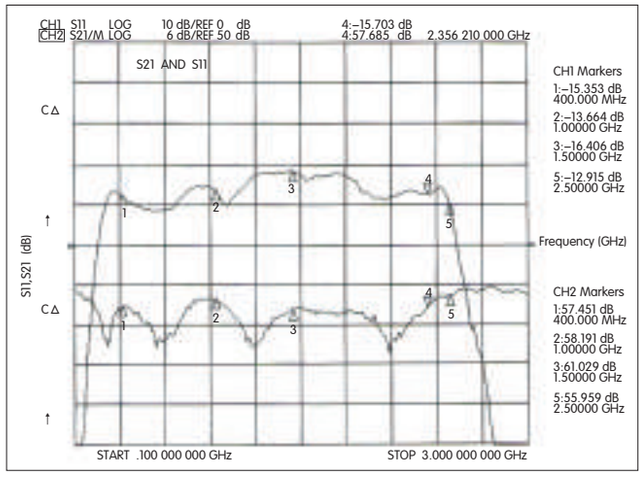
Measured S21 and S11 are shown in Fig.8. The power amplifier exhibits extremely flat small signal gain over a 5:1 fmax:fmin bandwidth. The average gain is about 57 dB and its ripple is +/− 3dB over the entire 0.5 G to



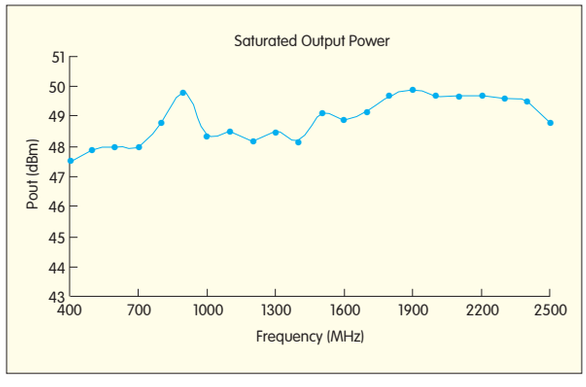
▲Figure 6. Power combining with 90 degree hybrid.



◀Figure 7. Broadband GaN amplifier with 90 degree hybrids.



◀Figure 8. Measured S21 and S11.



▲Figure 9. Measured output power from the GaN amplifier.

2.5 GHz bandwidth. The bandwidth can be further broadened when a hybrid with broader bandwidth is selected.

As shown in Fig. 9, more than 60 W of output power is achieved from this broadband amplifier. The output power is further increased if more transistors or transistors with higher power capability are used in the output stage. The operating frequency is also extended to 2600 MHz to cover the LTE frequency.

This GaN amplifier can cover multiple bands with a single amplifier. However, the linearity of the broadband amplifier is compromised to achieve output power covering wide bandwidth. The GaN amplifier can be used in the back–off mode to meet the linearity requirement of base stations and repeaters. The trade–off leads to low efficiency of the power amplifier. To maintain efficiency, the linearity is improved for a sub–band with proper tuning but not for the entire bandwidth. Therefore, additional linearization techniques such as envelop tracking, are also necessary if both high efficiency and high linearity are required when the amplifier is used for a unified broadband base station.

4 Conclusions

A broadband GaN high power amplifier is discussed in this paper. Very high output power can be achieved to cover multistandard and multiband applications with new GaN transistors. The GaN amplifier design can be improved to cover the full band from 700 MHz to 2600 MHz. A GaN amplifier can be operated in back–off mode to meet the requirements of linearity. It can also be tailored into several sub–bands to achieve better efficiency for each sub–band. Additional linearization techniques such as envelop tracking or DPD are necessary to further improve linearity and maintain high efficiency for the entire bandwidth [3]. This allows wireless infrastructure vendors to develop a single, highly–efficient multimode, broadband RF front end that can be deployed to meet various transmission standards anywhere in the world.

References

[1] D.M. Pozar, *Microwave Engineering*, 3 ed., New York: Wiley, 2004.
[2] S.C. Cripps, *RF Power Amplifiers for Wireless Communications*, Norwell, MA: Artech House, 1999.
[3] S.C. Cripps, *Advanced Techniques in RF Power Amplifier Design*, Norwell, MA: Artech House, 2002.

Biographies

Pengcheng Jia received his B.S. degree in electronics science and information systems from Nankai University, Tianjin, in 1995. He received his M.S. degree in electronic engineering from Tsinghua University, Beijing, in 1998, and his Ph.D. degree in electrical and computer engineering from the University of California, Santa Barbara, in 2002. His research on UCSB concerns the development of waveguide–based broadband high power spatial power combiner.

Dr. Jia joined ZTE as a micorwave system architect in December 2010. Prior to working with ZTE, Dr. Jia was the CTO of CAP Wireless at Newbury Park, California. He has worked with many microwave and mmwave systems, specifically focusing on broadband power amplifier design. He has pioneered coaxial waveguide–based spatial power combining technology and has successfully developed 2 G to 20 G and 20 G to 40 G ultra broadband power amplifier platforms. He has also developed many broadband power amplifiers using LDMOS and GaN transistors. Dr. Jia is a senior member of IEEE.

Single Mode DR Filters for Wireless Base Stations

Ji–Fuh Liang, Guo–Chun Liang, Marco Song, George He, and Tony An
(Pivotone Communication Technologies)

Abstract: This paper presents state–of–the–art high Q single–mode dielectric resonator (DR) cavity filters for PCS wireless base stations. DR cavity filters shrink the cavity size significantly more than waveguide cavity filters and offer about twice higher Q than coaxial resonators. Thus, they have important applications in wireless base stations operating below 2.5 GHz. Dual–mode and triple–mode DR cavity filters have existed for a while; however, single–mode DR cavity filters are predominant because they are cheaper to manufacture. This paper summarizes the main characteristics of TE01 mode DR cavities, including mode chart and field distribution, and compares cavity Q with waveguide and combline (coaxial) cavities. Dielectric combline and TM010 mode DR cavities are analyzed and compared to TE01 mode DR cavities. General filter design techniques are discussed, and several design examples are given to show how filter technology has developed.

Keywords: microwave resonator; cavity filter; DR cavity; dielectric resonator loaded cavity

1 Introduction

Dielectric resonator (DR) cavity filters have been used for satellite communications since the early 1980s because of their high Q (>10,000) and compactness. Temperature stability [1]–[3] and HEH11 dual mode are regarded as the major breakthroughs in DR cavity filter technology [4],[5]. In the early days, single TE01 mode cavities did not attract much attention for satellite applications because they provided no significant advantage over air–filled cylindrical dual–mode cavities [6] if transmission zeros could not be implemented in the stop band.

A TE01 mode filter with planar layout offers many advantages over an in–line configuration. The performance requirements for filters and multiplexer networks of wireless base stations and satellite applications are quite different. The cost of each filter and the issue of mass production are much more important than volume and weight in wireless base station applications. In the authors' opinions, the electrical performance of a state–of–the–art TE01 mode DR filter almost matches the performance of a HE11 dual–mode DR filter because the cross–coupling techniques have been developed for quasi–elliptic function filters. Also, asymmetric filter response with multiple transmission zeros in the stop band have been successfully implemented. TE01 single–mode filters have simple design, flexible layout options, and are cheaper than HE11 dual–mode filters to manufacture. However, they are bigger and heavier. Much literature on the TE01 mode DR cavity filter for wireless base stations can be found in the public domain [7]–[10].

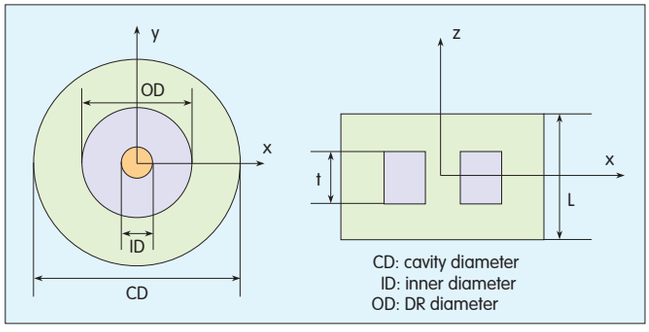
Another three single–mode DR cavity filters have emerged in the last decade [11]–[14]: half–wave TM mode, dielectric combline, and TM01 mode with both ends shorted. The half–wave has a smaller footprint; the dielectric combline is a pretty good solution in the middle range of cavity Q; and the TM01 mode with both ends shorted (denoted as TM010) has excellent volume efficiency to deliver cavity Q. However, the difficulty of shortening both ends of a dielectric rod limits the applications of TM010.

This paper summarizes the technological innovation of single–mode DR cavity filters for wireless base station applications. Section 2 summarizes aspects of cavity performance and design. Section 3 summarizes configurations with a variety of cross–coupling schemes that can be implemented in single–mode DR cavity filters. Section 4 presents several design examples. Section 5 concludes the paper.

2 Cavity Design and Performance

2.1 TE01 Mode Cavity

For high Q microwave filters, cavity electrical performance, and size and weight should be assessed simultaneously. This is because high Q microwave filters always occupy a significant amount of space in a



◀Figure1. Basic configuration of a TE01 mode DR cavity.

▼Table 1. Typical ceramic material from a commercial company

	E'	Q × F	Relative Cost/Volume	Linearity
8300	36	41,000	Lowest	Excellent
4300	43	43,000	Low	Good
4500	45	41,000	Low	Ok for Small Temp Range
3500	35	70,000	Moderate	Excellent
8700	30	100,000	High	Good
2900	30	110,000	High	Good

transceiver subsystem, especially in L–band.

Fig. 1 shows a basic configuration of a TE01 mode in a DR cavity. The conductive enclosure can be a circular or rectangular cavity. In order to limit the loss from the conductive enclosure, the cavity diameter (CD) is usually greater than 1.5 times the DR diameter (OD), and the height of the cavity (L) is about 3 times that of DR thickness (t).

Typical high Q dielectric material that is commercially available is listed in Table 1. Material with a dielectric constant of 30 may yield the highest dielectric Q. However, it is very expensive and not practical for most applications below 2.5 GHz.

The design of a TE01 mode DR cavity should take into account cavity Q_u, size, and spurious responses simultaneously. These can be computed using a rigorous radial mode–matching technique [15], [16] or with generalized 3–D EM wave simulators such as HFSS or CTS. Cavity Q_u, size, and spurious responses are dictated by the DR aspect ratio, which is defined as the ratio of DR diameter (OD) to DR thickness (t) (Fig. 1). The aspect ratio of the DR cavity should be properly chosen; otherwise, the high–order modes may be too close to the working mode. Mode charts [7], [15]–[17] have been proposed for the design of DR cavities. The mode chart of a solid DR of 1.9 GHz TE01 mode with dielectric constant of 44 is shown in Fig. 2 (a). It is also well known that opening a hole in the center of TE01 mode DR can increase the spurious–free region of the cavity (Fig. 2 (b)). This is because the TE01 mode DR cavity has a minimum electric field at the DR center, and all other closer spurious modes have a maximum electric field. The results in Fig. 2 suggest that the aspect ratio of the TE01 mode DR (ε_r = 44.0) cavity is around 2.5, and the diameter of the center hole can be opened up to 35% of the DR diameter. The relative mode locations in the frequency spectrum are not a strong function of the conductive enclosure, except in the case of TM mode. The mode charts in Fig. 2 [7] suggest that HE11 (HEH11) and HE12 (HEE11) modes are the closer spurious frequencies, but TM01 mode is usually the one that causes interference in pass band if the cavity is not tuned properly. There are three characters and two numbers used for the index of the DR cavity mode. The first two characters comprise TE, TM or HE—which represent transverse E–field, transverse H–field, and hybrid mode. The first number represents the azimuth variation, and the second number represents the order of the modes in the frequency domain according to the condition defined by the previous

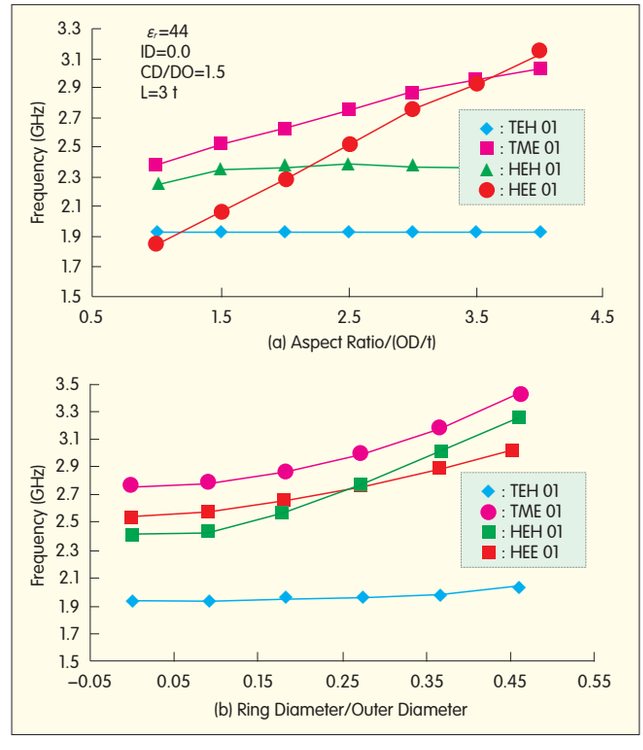
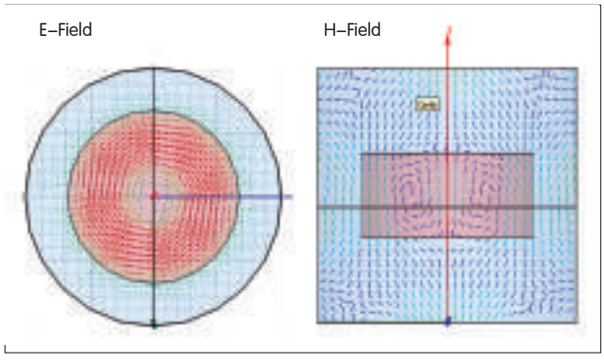
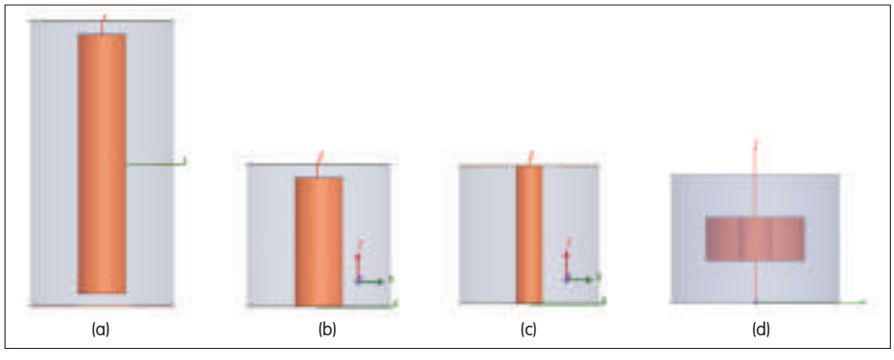


Figure 2. ▶ Mode chart of a TE01 mode solid and ring resonator (ε_r = 44.0).

Ji–Fuh Liang, Guo–Chun Liang, Marco Song, George He, and Tony An



◀Figure 3. E–field (on the x–y plane) and H–field (on the x–z plane) of a TE01 mode DR cavity.



▲Figure 4. Three possible configurations of TM01 mode (a, b, and c), and a TE01 mode DR cavities (d).

characters and number. A typical E–field (on the x–y plane) and H–field (on the x–z plane) of TE01 mode are computed by HFSS, as shown in Fig. 3. These are important for input/output and inter–cavity design for a DR cavity filter.

2.2 Dielectric Combline and TM Mode Cavity

The high Q TE01 mode dielectric resonator cavity usually has an aspect ratio of the DR puck range from 2.0 to 3.5 to open up a spurious–free region for TEH01 or HEH01 mode. This aspect ratio ranges from 0.15 to 0.5 for medium Q TM mode, and it looks like a rod rather than a disk. Another point of view is that the dielectric rod plays a similar role to a metal rod in a coaxial resonator. So the basic properties of the TM01 mode, such as field distribution and cavity Q, is similar to a coaxial resonator except that Q degradation in the dielectric rod, which is dominated by dielectric loss, is much smaller than in the metal rod. Also, the spurious performance can be quite different.

There are three operating conditions for the DR TM mode filter: with half–wave resonators, quarter–wave resonators, and with both sides of the dielectric rod shorted—TM010 mode (the third number denotes no field variation in the z–direction), as shown in Fig. 4(a), (b) and (c). The characteristics of a 2 GHz cavity configured as shown in Fig. 4(a), (b), and (c) are designed by simulation using HFSS. The performance results using TE01 mode, dual mode HE11 mode, and metallic coaxial resonator are also included for comparison (Table 2). The dielectric constant of the ceramic puck and rod is taken as 45.0, and the loss tangent is 4e–5 (i.e. dielectric Q, Q_d=25,000). For the results in Table 2, the metal is assumed to be silver plated. Fig. 5 (a) and (b) show the field distributions of the TM mode cavities; Fig. 5 (c) and (d) are the magnitudes of E– and H–fields of the x–z plane of the dielectric combline; and Fig. 5 (e), (f) are the TM010 mode. Dielectric combline has a field variation along the z–axis (Fig. 5 (c) and (d)) and its length is pretty much fixed by resonant frequency. The TM010 mode has a uniform field distribution along the z–axis (Fig. 5 (e) and (f)), which means that cavity high can be reduced without affecting resonant frequency. This provides an additional dimension to the trade–off between cavity Q and size.

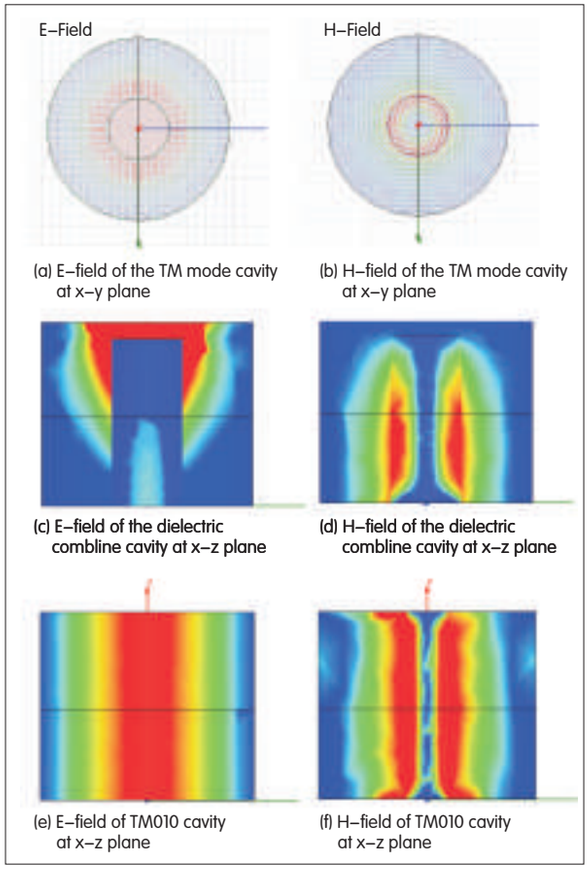
The half–wave resonator provides high Q, which can be as high as in TE01 mode. However, the cavity is much longer than in TE01 mode. Although the half–wave resonator cavity has a smaller footprint than the TE01 cavity, its overall size is still bigger. The dielectric combline works as a quarter–wave resonator, creating smaller volume and medium–range Q for the DR cavity technology. The TM01 mode resonator with both ends shorted has lower Q but smaller DR volume. For this reason, the cost of ceramic material is much cheaper. When the TM01 mode resonator is the same size as the metallic combline resonator, the metallic coaxial resonator has the smallest cavity Q.

Table 2 shows high Q design of different modes. With variations in size

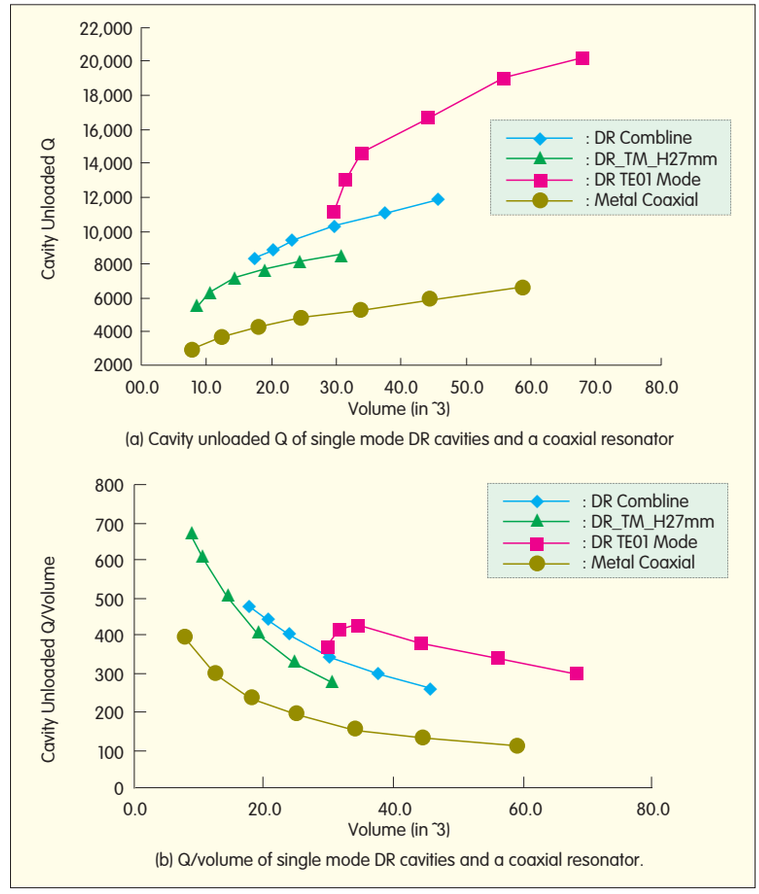
▼Table 2. Cavity Q, volume of single mode DR cavity

	Cavity		DR Disk/Post		Cavity Performance				DR
Design	D (mm)	H (mm)	D (mm)	H (mm)	fo (MHz)	Q_u	Volumn (in cm^3)	Q_u/Vol.	Volume
A	46.0	30	34.5	13.0	2023	19,776	49.8	793.7	12.15
B	45.9	30	27.0	10.0	2022	18,487	49.6	372.6	5.15
C	38.0	66	12.7	60.0	2026	18,410	74.8	246.1	7.6
D	38.0	33	12.7	30.0	2026	11,282	37.4	301.6	3.8
E	38.0	33	7.0	33.0	2027	9419	37.4	251.8	1.27
F	38.0	33	12.7	27.1	2026	5531	37.4	147.9	0.00

A: DR HE11 dual–mode cavity
B: TE01 mode
C: Half–wave TM mode
D: Dielectric combline
E: TM mode with both ends shortened
F: Coaxial cavity



▲Figure 5. Field distributions of the TM mode, dielectric combline, and TM010 mode cavities.



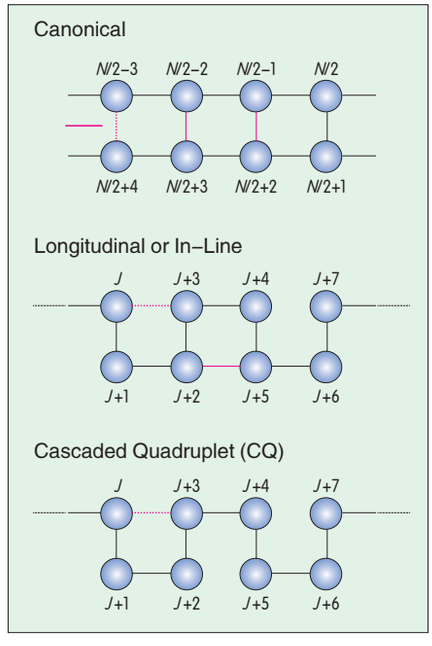
▲Figure 6. The results of high Q design of different modes.

for each design, the results are shown in Fig. 6. TE01 mode has better Q and volume efficiency than other modes, and it is suitable for high Q applications (cavity > 10,000). Between 8000 and 10,000, TE01 efficiency drops significantly; thus, TE01 is not practical. Because it has a smaller than TE01, and less ceramic material is needed, dielectric combline fills the gap in this range of applications. Below a cavity Q of 8000, a dielectric combline cavity becomes impractical because the diameter of the dielectric rod needs to be significantly increased. A bigger puck volume means higher cost. Both–ends shorted TM010 mode also needs to be used. Between cavity Q of 4000 and 8000, TM010 mode DR cavity provides better volume efficiency than a metal coaxial resonator (Fig. 6). Q_u and Q_u/volume of the combline and TM010 mode DR cavity in Fig. 6 form a continuous performance spectrum as a function of volume and cavity Q.

3 Filter Topologies with A Variety of Cross–Coupling Schemes

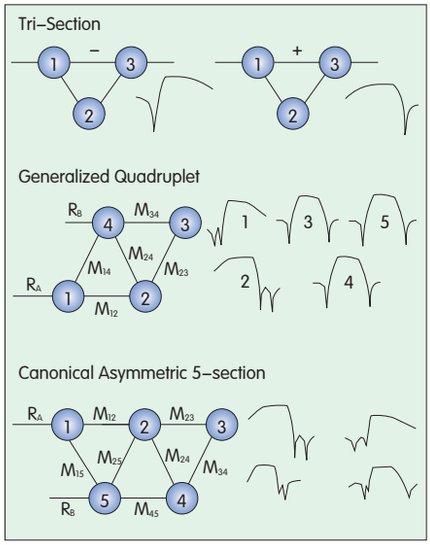
There is a variety of cross–coupling schemes for microwave filter design that can be used to meet the filter rejection requirement with reduced filter order. The available forms for symmetric responses are: canonical [16]–[18], longitudinal [19], and cascaded quadruplet [20] (Fig.7). For asymmetric responses, cascaded tri–sections [21],[22] and cascaded canonical asymmetric building blocks [23] can be used. The building blocks of asymmetric filter response are shown in Fig. 8.

Using the building blocks in Fig. 8 for low order filters and cascading two of them for higher order filters (Fig. 9) provides microwave filter designers

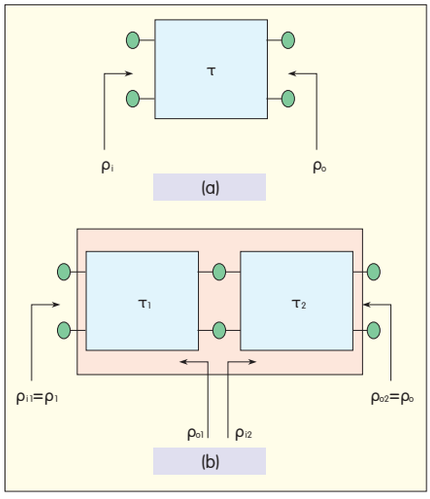


▲Figure 7. Cross–coupling schemes for symmetric filter responses.

Ji–Fuh Liang, Guo–Chun Liang, Marco Song, George He, and Tony An



▲Figure 8. Building blocks for asymmetric filter responses.



▲Figure 9. Principle of cascading the building blocks for asymmetric filter.

▼Table 3. Implementation of symmetric zeros with different filter topologies

Filter Order	Canonical	Longitudinal	Proposed
6	4	2	2 (1,1)
7	4	2	4 (2,2)
8	6	4	4 (2,2)
9	6	4	6 (3,3)

▼Table 4. Implementation of asymmetric zeros with different filter topologies

Filter Order	Canonical	Tri–sections	Proposed
6	4	2	3 {(2,1), (1,2), (3,0), (0,3)}
7	5	3	4 {(3,1), (1,3), (2,2)}
8	6	3	5 {(3,2), (2,3), (2,2)}
9	7	4	6 {(3,2), (2,3), (3,3)}

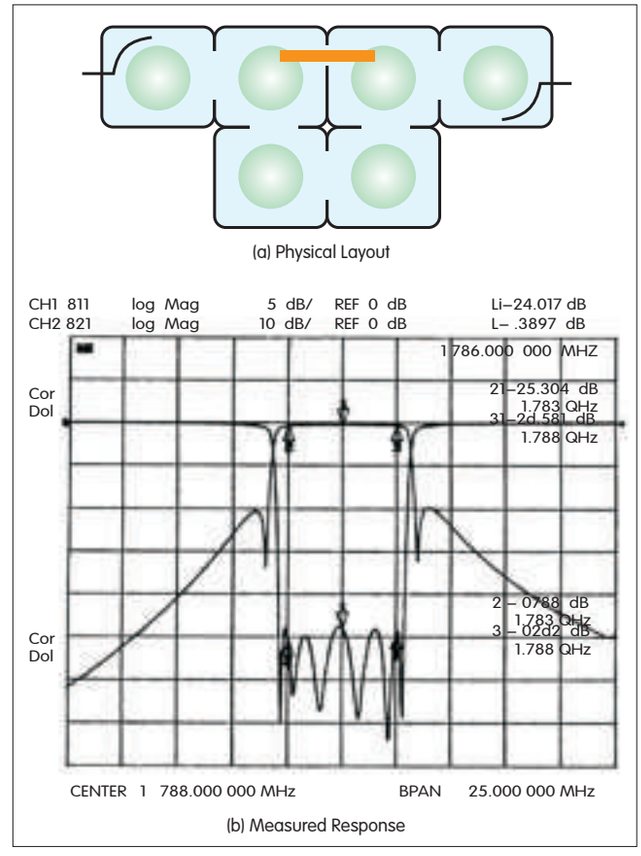


Figure 10. ▶ The physical layout and measured results of a 6–pole TE01 mode DR cavity filter.

with alternatives for filter topology and implementation. In Tables 3 and 4, the maximum transmission zeros that can be implemented with this approach are compared with other options and summarized for symmetric and asymmetric filter responses, respectively. The realizable transmission zeros of the approach in Fig. 9 is two less than or the same as the canonical approach for symmetric filter response. It is one less than the canonical approach for asymmetric filters. However, in most filter applications, it is not practical to implement the maximum zeros. As well as having an advantage over longitudinal and cascaded tri–section in terms of possible zeros, the approach in Fig. 9 also offers a tuning mechanism for independent placement of transmission zeros.

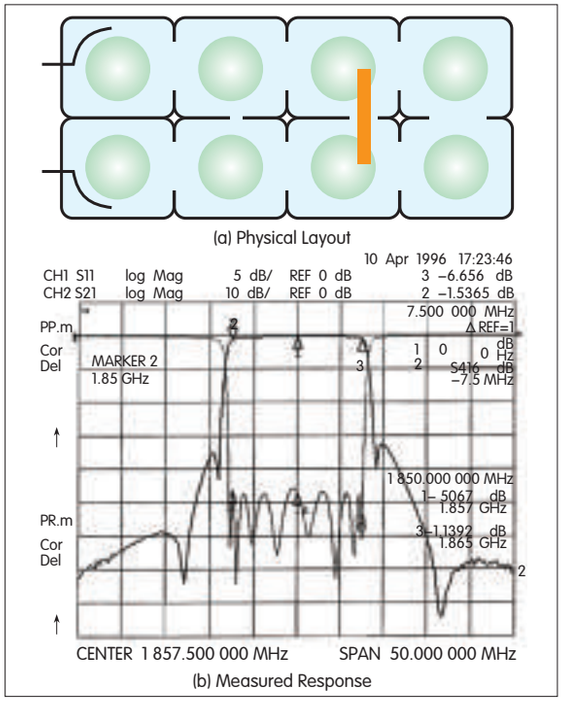
Design Examples

(1) Example 1: 6–pole and 8–pole quasi–elliptic–function filters

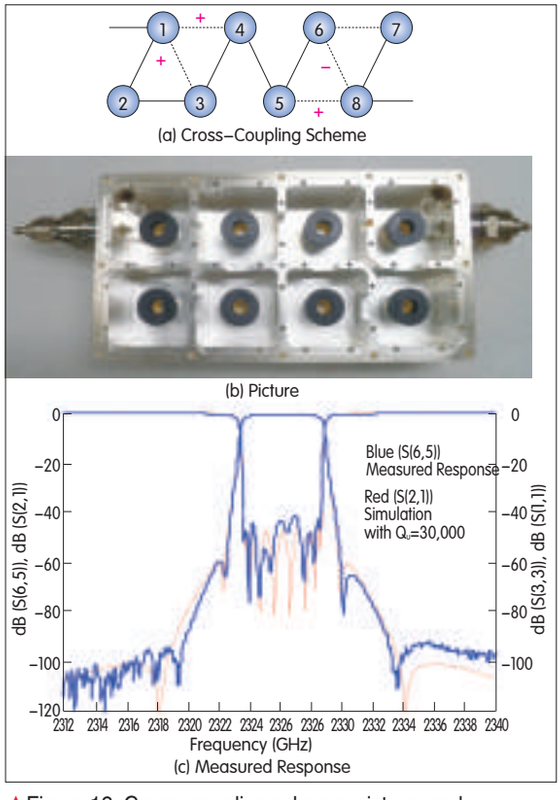
The physical layout and measured results of a 6–pole (5 MHz bandwidth) and 8–pole (15 MHz bandwidth) quasi–elliptic–function filter at PCS frequency are shown in Fig. 10 and Fig. 11. The 6–pole filter has one 2–5 cross coupling, and the 8–pole filter has 3–6 and 2–7 cross couplings. The 6–pole filter is realized by high Q DR puck with dielectric constant of 29, and the effective–filter unloaded Q is 24,500.

(2) Example 2: Very high Q (Q_u =29,000), 8–pole quasi–elliptic–function filter

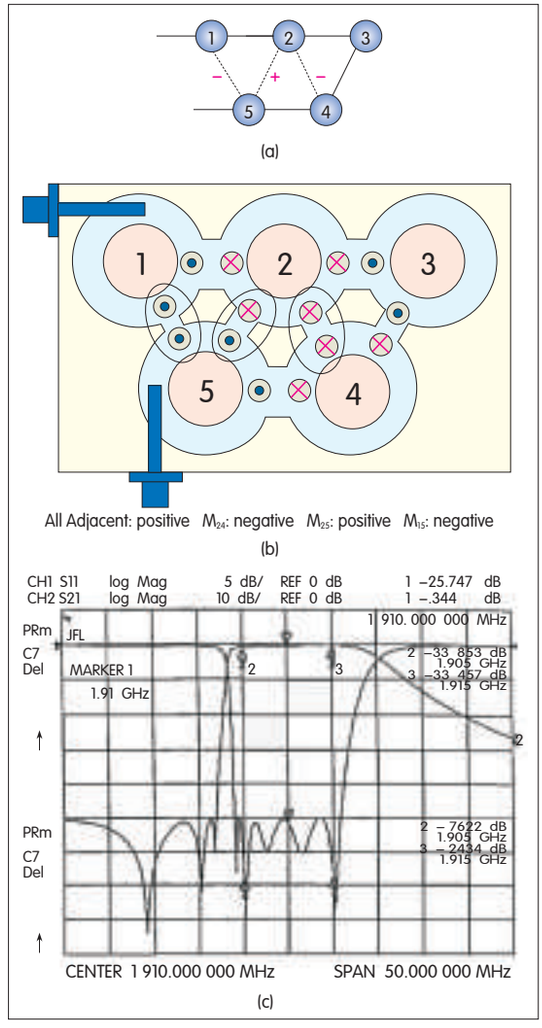
In this example, an 8–pole filter with two transmission zeros in each side of the stop band is implemented by cascading two generalized quadruplets. Compared with the symmetric canonical filter in Fig. 11, this filter better balances the transmission zeros on both sides of stop band. The schematic, layout, and measured



▲Figure 11. The physical layout and measured results of an 8–pole TE01 mode DR cavity filter.



▲Figure 12. Cross–coupling scheme, picture, and measured response of a 8–pole TE01 mode DR cavity filter with cavity Q of 29,000.



◀Figure 13. Filter cross–coupling scheme, physical layout, and measured response of a 5–pole TE01 mode DR cavity filter.

response of the 8–pole filter are shown in Fig. 12.

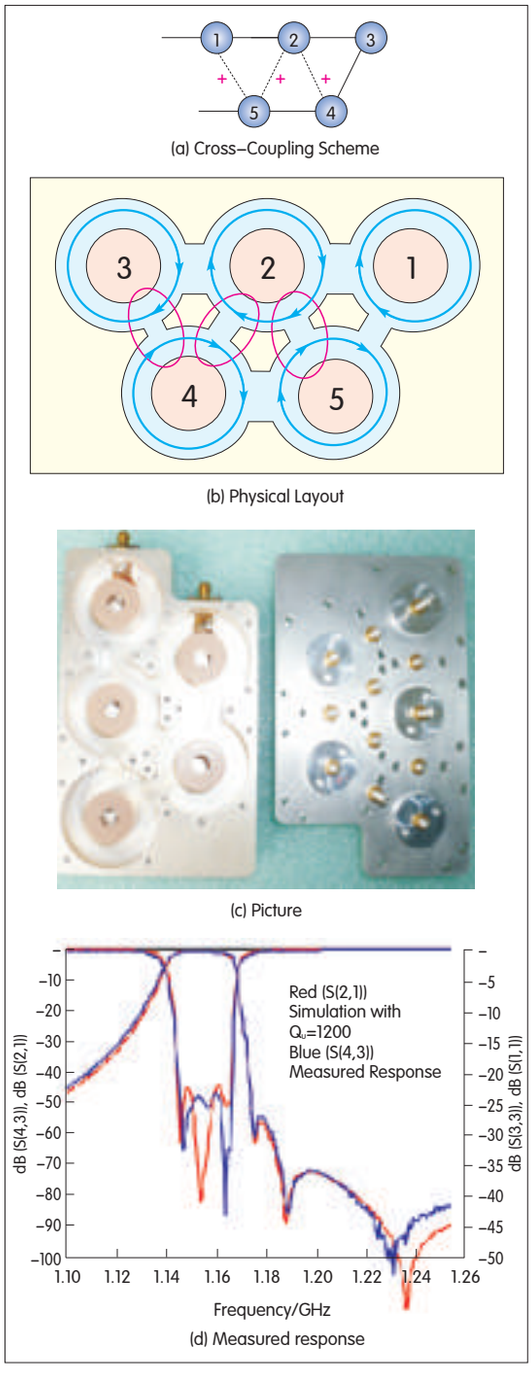
(3) Example 3: 5–pole TE01 mode canonical asymmetric filter

The TE01 mode DR cavity is a very interesting microwave filter technology, especially because of its asymmetric filter responses. In Fig. 13, the coupled magnetic field runs perpendicular to the plane of the page when looking at the top view of the planar mechanical layout. More physical detail can be seen in the modal field distribution in Fig. 3. For the cross–coupling scheme of the 5–pole filter in Fig. 13(a), the physical layout and orientation of the magnetic field of the cavity sidewall are shown in Fig. 13 (b). M_{24} (–), M_{25} (+) and M_{15} (–) are implemented by coupling irises. The measured results are shown in Fig. 13(c). Two negative cross–couplings are implemented by an iris that is inductive. Detailed description and analysis of this filter can be found in [7] and [23].

(4) Example 4: A 5–pole TM mode canonical asymmetric filter

Compared with a TE01 mode DR cavity, the TM mode's filter has a quite different characteristic. For the cross–coupling scheme in a 5–pole filter with cross–couplings M_{24}, M_{25} and M_{15} (Fig. 14(a)), the couplings are also implemented by an iris. But they are all positive, which yields three transmission zeros on the high–side

Ji–Fuh Liang, Guo–Chun Liang, Marco Song, George He, and Tony An



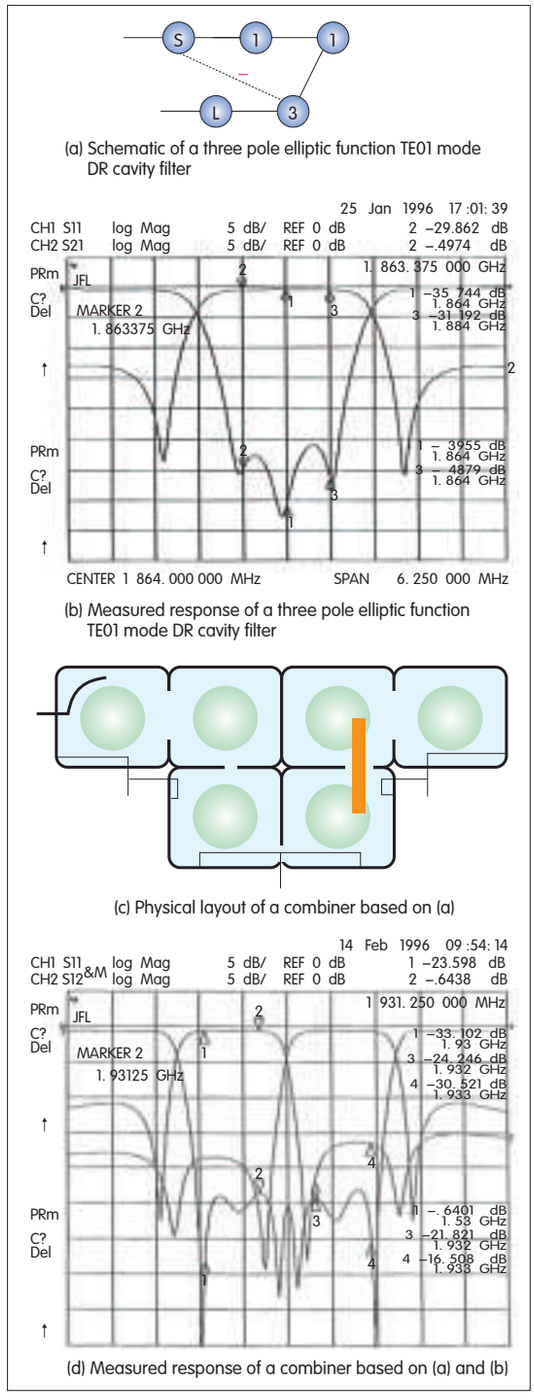
(a) Cross–Coupling Scheme

(b) Physical Layout

(c) Picture

(d) Measured response

▲Figure 14. Cross–coupling scheme, physical layout, picture, and measured response of a 5–pole TE01 mode DR cavity filter.



(a) Schematic of a three pole elliptic function TE01 mode DR cavity filter

(b) Measured response of a three pole elliptic function TE01 mode DR cavity filter

(c) Physical layout of a combiner based on (a)

(d) Measured response of a combiner based on (a) and (b)

▲Figure 15. 3–pole elliptic function TE01 mode DR cavity filter and combiner.

stop band. The picture of the filter is Fig. 14(c) and the measured response is shown in Fig. 14(d).

(5) Example 5: 3–pole elliptic function TE01 mode DR cavity filter and combiner

A true odd–order elliptic function filter requires non–adjacent coupling between the source or load and an internal resonator. The schematic and measured response of a 3–pole elliptic function filter is shown in Figs. 15(a) and (b). The physical layout and measured response of a combiner based on (a) is shown in Figs. 15(c) and (d). The non–adjacent coupling M_{03} and input coupling RA are realized by inductive loops to construct proper phase between cavity number one and three for positive cross–coupling.

5 Conclusions

This paper reviews state–of–the–art single–mode DR cavity filters for wireless base stations. The cavity characteristics of three TM modes at 2 GHz are analyzed using HFSS. The design of DR cavity filter for high, medium, and low Q operation is compared with the design for TE01 DR cavity modes. TE01 mode is suitable for high Q (>12,000) applications, dielectric combline is suitable for medium Q (8000–12,000) applications, and TM010 mode is suitable for lower Q (<8000) applications. A variety of cross–coupling schemes for implementing symmetric and asymmetric transmission zeros are presented, and topology based on cascading canonical asymmetric building blocks is discussed.

Designs with excellent measured performance are presented, with 6 and 8–pole quasi–elliptic–function filters taken as examples. An 8–pole filter with two transmission zeros on each stop band is an example of how, by cascading

asymmetric building blocks, symmetric transmission zeros can be better balanced. This kind of 8–pole filter is high Q (around 30,000), and represents state–of–the–art high Q dielectric material technology. A 5–pole DR TE01 mode canonical asymmetric filter is an example of how three low–side transmission zeros can be implemented by three non–adjacent coupling irises. For a DR TM010 mode cavity filter, three non–adjacent coupling irises can yield three high–side transmission zeros. The combining of a DR TM010 mode cavity filter with a 3–pole elliptic function filter highlights the progress that has been made in single mode DR cavity filters for wireless base stations.

References

[1] D. J. Passe and R. A. Pucel, "A temperature–stable bandpass filter using dielectric resonators," *Proc. IEEE*, vol. 60, no. 6, pp. 730, Jun. 1972.
[2] K. Wakino, T. Nishikawa, S. Tamura, and Y. Ishikawa, "Microwave bandpass filters containing dielectric resonators with improved temperature stability and spurious response," in *IEEE Int. Microw. Symp. Digest*, Palo Alto, CA, 1975, pp. 63–65.
[3] J. K. Plourde and D. F. Linn, "Microwave dielectric resonator filters utilizing Ba2Ti9O20 ceramics," in *IEEE Int. Microw. Symp. Digest*, San Diego, CA, 1977, pp. 290–293.
[4] S. J. Fiedziuszko, "Dual–mode dielectric resonator loaded cavity filters," *IEEE Trans. Microw. Theory Tech.*, vol. 30, no. 9, pp. 1311–1316, Sep. 1982.
[5] S. J. Fiedziuszko, "Engine–block dual mode dielectric resonator loaded cavity filter with non–adjacent couplings," in *IEEE Int. Microw. Symp. Digest*, San Fransisco, CA, 1984, pp. 285.
[6] Kudsia, R. Cameron and W. C. Tang, "Innovations in microwave filters and multiplexing networks for communications satellite systems," *IEEE Trans. Microw. Theory Tech.*, vol. 40, no. 6, pp. 1133–1149, Jun. 1992.
[7] Ji–Fuh Liang and William D. Blair, "High Q TE01 mde DR filters for PCS wireless base stations," *IEEE Trans. Microw. Theory Tech.*, vol. 46, no. 12, pp. 2493–2500, Dec. 1998.
[8] R. R. Mansour, "Filter technologies for wireless base stations," *IEEE Microwave Mag.*, vol. 5, no. 1, pp. 68–74, Mar. 2004.
[9] C. Wang and K. A. Zaki, "Dielectric resonators and filters," *IEEE Microw. Mag.*, vol. 8, no. 5, pp. 115–127, Oct. 2007.
[10] Jorge A. Ruiz–Cruz, C. Wang and K. Zaki, "Advances in microwave filter design techniques," *Microwave Journal*, Nov. 2008.
[11] C. Wang, K. A. Zaki, A. E. Atia, and T. G. Dolan, "Dielectric combline resonators and filters," *IEEE Trans. on Microw. Theory and Tech.*, vol. 46, no. 12, pp. 2501–2506, Dec. 1998.
[12] Toshio Ishizaki, "Temperature–stable dielectric TM010–mode resonator and its application to compact base station filter," *IEICE Electronic Express*, vol.7, no.6, pp. 454–459, 2010
[13] M. Hoft, "Bandpass filter using TM–mode dielectric rod resonators with novel input coupling," in *IEEE MTT–S Int. Microw. Symp. Digest*, Boston, MA, 2009, pp. 1601–1604.
[14] Ming Yu, D. Smith, and M. Ismail, "Half–wave dielectric rod resonator filter," in *IEEE Int. Microw. Symp. Digest*, Fort Worth, TX, 2004, pp. 619–622.
[15] S. W. Chen and K. A. Zaki, "Dielectric ring resonators loaded in waveguide and on substrate," *IEEE Trans. Microw. Theory Tech.*, vol. 39, no. 12, pp. 2069, Dec. 1991.
[16] Xiang–peng Liang and K. A. Zaki, "Modeling of cylindrical dielectric resonators in rectangular waveguides and cavities," *IEEE Trans. Microw. Theory Tech.*, vol. 41, no. 12, pp. 2174–2181, Dec. 1993.
[17] A. E. Atia and A.E. William, "New types of bandpass filters for satellite transponders," *Comsat Tech. Rev.*, vol. 1, pp. 21–43, Fall 1971.
[18] A. E. Atia and A.E. William, "Narrow bandpass waveguide filters," *IEEE Trans. Microw. Theory Tech.*, vol. 20, no. 4, pp. 258–265, Apr. 1972.
[19] R. J. Cameron, "General prototype network synthesis methods for microwave filters," *ESA Journal*, vol. 6, pp. 193–206, 1982
[20] R. Levy, "Direct synthesis of cascaded quadruplet (CQ) filters," *IEEE Trans. Microw. Theory Tech.*, vol. 43, no. 12, pp. 2940–2945, Dec. 1995.
[21] H. C. Bell, Jr, "Canonical Asymmetric coupled–resonator filters," *IEEE Trans. Microw. Theory Tech.*, vol. 30, no. 9, pp. 1335–1340, Sep. 1982.
[22] J. D. Rhode and R. J. Cameron, "General extracted pole synthesis technique with application to low–loss TE01 mode filters," *IEEE on Microw. Theory and Tech.*, vol. 28, no. 9, pp. 1018–1027, Sep. 1980.
[23] Ji–Fuh Liang and Dawei Zhang, "General coupled resonator filters design based on canonical asymmetric building blocks," in *IEEE MTT–S Int. Microw. Symp. Digest*, Anaheim, CA, 1999, pp. 907–910.

Biographies

Ji–Fuh Liang (Jifuh.liang@pivotone.com)received the B.S. Degree in electronics engineering from National Chaio–Tung University, Taiwan in 1981, M.S. Degree in electrical engineering from National Taiwan University, Taiwan in 1985, and Ph.D. degree in electrical engineering from University of Maryland in 1994.

From 1985 to 1988 he was a member of technical staff and Project Leader at Microelectronics Technology Inc., Hsin–Chun, Taiwan. He maybe the earliest microwave engineer using dielectric resonator cavity for PCS 5MHz and 15MHz filter and HTS development. He joint BenQ corporation from 2000 to 2005 in Taiwan, as senior member of technical staff, senior manager and director for CDMA and handset technologies.

He involved in FBAR filter design from 2006 to 2010 at Avago, where he designed first FBAR Wi–Fi and WiMAX filters for wireless handheld devices. Since March, 2010, he joint Pivotone Communication Technologies as CTO and VP of technology. His current research areas include DR cavity filter, waveguide filter and multiplexer for wireless base stations, and point–to–point communications.

Guo–Chun Liang (gcliang@pivotone.com) received the B.S. degree from the East China Institute of Technology, China in 1982, the master's degree from the University of Electronics Science and Technology of China (UESTC) in 1985, and Ph.D. degree from University of California at Berkely in 1990. All majors were in electrical engineering.

He worked at the center of UESTC in 1985–1986, and developed a series of RF and microwave devices there. He had been with Conductus, Inc., Sunnyvale, CA, from 1990 to 2001, where he developed most advanced superconductor components and filters for wireless base station applications. In 2002, he found Allizon Communication Technology in Shanghai, which was acquired by Smith Group. He is now the president and CEO of Pivotone Communication Technology Inc.

Marco Song (marco_song@pivotone.com) received the B.S. degrees in microwave from University of Electronics Science and Technology of China (UESTC) in 1999. He is now the RF manager with Pivotone Communication Tech., Inc. His research interests include RF and microwave passive components and measurements.

George He (george_he@pivotone.com) received the B.S. degrees in microwave from University of Electronics Science and Technology of China (UESTC) in 2001. He joint Shanghai Institute of Control and Communication, from 2001 to 2003, for microstrip circuits research and development. In 2003, he joint Metac communication equipment inc as a senior RF engineer, and was responsible for developing coaxial resonator, dielectric cavity filters and LNA for wireless base station applications. In 2009, he joint Powerwave and continued his research and development in similar applications. Now, he is a senior RF engineer and project leader in Pivotone Communication Technologies, Inc.

Tony An (tony_an@pivotone.com) graduated from University of Electronic Science and Technology of China (UESTC), majored in Electromagnetic Fields and Microwave Engineering (2001.9–2005.7). Tony has been engaged in research and development of passive radio frequency device, include GSM, CDMA, MAX, LTE, waveguide filters, diplexers and other radio frequency devices. His experience of product design ranges from 30 MHz–42 GHz. He also has deep understanding and good experience in designing other passive products. Now, Tony is working with Pivotone as RF senior engineer and project leader, mainly in waveguide filter, diplexer and other related components.

Design of a Magneto–Electric Dipole Element for Mobile Communication Base Station Antennas

Hang Wong[1] *and Kwai Man Luk*[2]
(1. State Key Laboratory of Millimeter Waves (Hong Kong);
2. Department of Electronic Engineering, City University of Hong Kong)

Abstract: The magneto–electric dipole antenna is a kind of complementary antenna composed of a planar electric dipole and a shorted patch antenna. It has excellent electrical characteristics including wide impedance bandwidth, low cross–polarization, low back lobe radiation, nearly identical E–plane and H–plane patterns, stable radiation pattern, and steady antenna gain over the operating frequency range. In this paper, the basic characteristics of a linearly polarized magneto–electric dipole antenna are reviewed, and a dual–polarized antenna element based on the magneto–electric dipole is presented. The design of a conical beam wideband antenna with horizontal polarization is also described. These antennas have practical applications in modern 2G, 3G, LTE, WiFi, and WiMax wireless communication systems.

Keywords: base station antenna; magneto electric dipole; wideband antenna; dual polarization

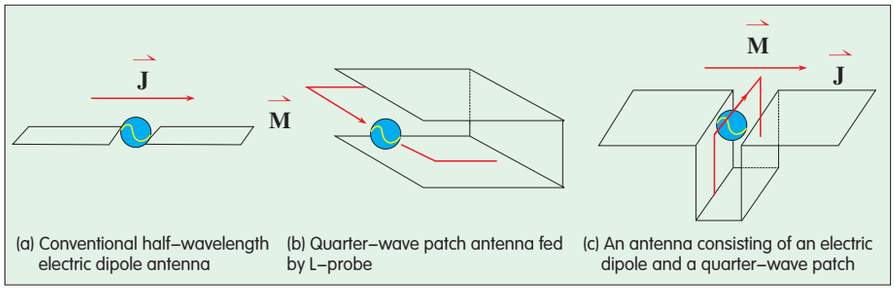
1 Introduction

2G, 3G, LTE, Wi–Fi, and WiMAX, demand wideband unidirectional antennas with low cross–polarization, low back radiation, symmetric radiation pattern, and stable gain over the operating frequency range. Antennas with these excellent electrical characteristics can save cost, space, and energy because several wireless communication systems can be accommodated. There are three conventional ways of implementing wideband low–profile antennas with a unidirectional radiation pattern. They are: (1) directed dipoles; (2) wideband patch antennas; and (3) complementary antennas.

Dipole antennas are commonly used in wireless communication systems because of their reasonably wide bandwidth, good radiation characteristics, and ease of construction. They are capable of a directional or bi–directional radiation pattern [1]–[3]. Much effort has been put into developing wideband dipole antennas because of the continuously expanding range of wireless services for voice and data. Methods such as flared dipole arms [4], bowtie shaped dipole [5]–[6], flat dipole [7] and the use of parasitic elements [8] have been proposed to achieve wideband for dipole antennas. Bandwidth of 30% to 100% can be achieved if a wideband balun is included. However, wideband directed dipole antennas cannot easily maintain a stable radiation pattern over the operating frequency range. Their radiation patterns vary substantially with frequency.

Another popular unidirectional antenna is the microstrip or patch antenna. There are several designs for wideband patch antennas including the L–probe patch antenna [9]–[13], the aperture coupled patch antenna [14]–[16], the stacked patch antenna [17]–[21] and the U–slot patch antenna [22]–[27]. These antennas have a wide impedance bandwidth ranging from 20% to 40%, which is sufficient for many modern wireless communication systems. However, these antennas [9]–[27] have drawbacks such as high cross–polarization, large variation in gain, and beamwidth over the operating frequency range. Although techniques such as anti–phase cancellation [28], twin–L probes coupled feed [29], and meandering–probe feed [30]–[34] can be employed to suppress cross polarization, the resulting antennas still have large gain and beamwidth variations with different frequencies. They also have large differences in E–plane and H–plane beamwidths.

To achieve stable radiation characteristics over the operating frequency range, a third approach is to develop a complementary antenna consisting of an electric dipole and a magnetic dipole. The technique of using a complementary antenna to achieve equal E–plane and H–plane patterns was revealed several decades



▲Figure 1. Principle of the design.

ago [35], [36]. It is well known that an electric dipole has a figure–8 radiation pattern in the E–plane and a figure–O pattern in the H–plane. A magnetic dipole has a figure–O pattern in the E–plane and a figure–8 in the H–plane. If both electric and magnetic dipoles can be excited simultaneously with appropriate amplitude and phase differences, a unidirectional radiation pattern with equal E–plane and H–plane can be obtained. A practical design shown in [37] consists of a passive dipole placed in front of a slot. This idea is based on a slot–and–dipole combination [38]–[40]. However, these designs [35]–[40] have either narrow bandwidth or bulky structure. They may not be able to fulfill the stringent requirements of modern wireless communication systems. Recently, a new wideband unidirectional antenna element comprising a planar dipole and a vertically–oriented shorted patch antenna was presented [41]. This antenna was developed based on a complementary concept. It has a simple structure, wide bandwidth, low cross–polarization, symmetrical radiation pattern, and very low back radiation. Because of low back radiation, the gain and beamwidth of the antenna do not vary significantly with frequency. As a result, the gain and efficiency of the antenna are higher than that of many other available antenna elements. The antenna has many applications in modern wireless communication systems.

2 Magneto–Electric Dipole Antennas

To implement a complementary antenna, an appropriate electric dipole and magnetic dipole needs to be selected. Among many choices of electric dipoles, a planar dipole antenna is chosen (Fig.1a) and a wideband short–circuited patch antenna is chosen as the magnetic dipole (Fig.1b) [41]. To combine these two antennas, a short–circuited patch is placed so that its open end can be connected to the horizontal planar electric dipole, as shown in Fig.1c [41]. Based on this method, a new wideband antenna is developed after a detailed parametric study, and its prototype is shown in Fig.2 [41]. This antenna operates at 2.5 GHz. The planar dipole has a width W=60 mm (0.5 λ) and a length L=30 mm (0.25 λ). The shorted patch antenna has a length H=30 mm (also close to 0.25 λ). The distance between the two vertical plates is S=17 mm, which is about 0.14 λ. The width of the dipole and the patch W is about 0.5 λ. The size of the ground plane can be changed to adjust the back radiation—an acceptable choice is 160 mm ×160 mm (1.3 λ ×1.3 λ).

An Γ–shaped probe is used as the feed. This consists of three parts made by folding a rectangular strip of metal. The first part is vertically oriented. One end is bonded to the coaxial launcher mounted underneath the ground plane. Together with the vertical plate of the shorted patch antenna, this part functions as an air microstrip line with a 50 Ω characteristic impedance. The line transmits the signal to the second part of the feed. The second part is placed horizontally to excite the planar dipole and the shorted patch antenna simultaneously. The input resistance of the antenna depends on the length of this part. It also carries an inductive reactance that can affect the matching performance of the antenna. Together with the second vertical plate, the third part of the feed forms an open–circuited microstrip line. By choosing a suitable length for this part, its equivalent capacitive reactance can be used to suppress the inductive reactance that is caused by the second part.

The radiation patterns of a thin dipole, a planar dipole, and a magneto–electric dipole (shown in Fig.3) are simulated [41]. They have the same ground plane size of 160 mm×160 mm and the same antenna height H=30 mm (0.25 λ). For these three cases, the length of the dipole is L= 60 mm (0.5 λ), referring to an operating frequency of 2.5 GHz. For the first two cases, each antenna needs a balun for excitation. For the third case, the antenna is simply excited by a Γ–shaped strip feed.

The thin dipole in Fig.3a and the planar dipole in Fig.3b have the same radiation pattern, that is, high back

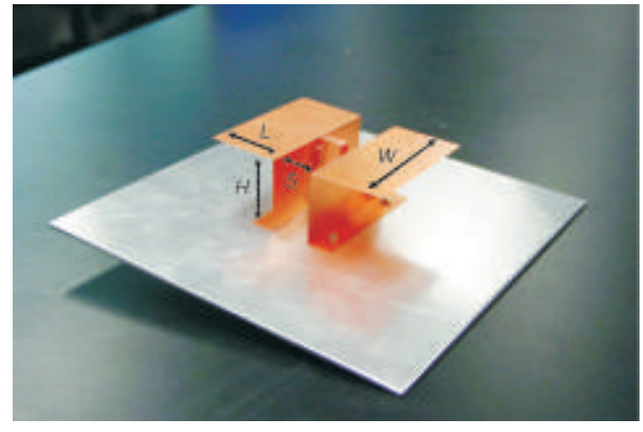
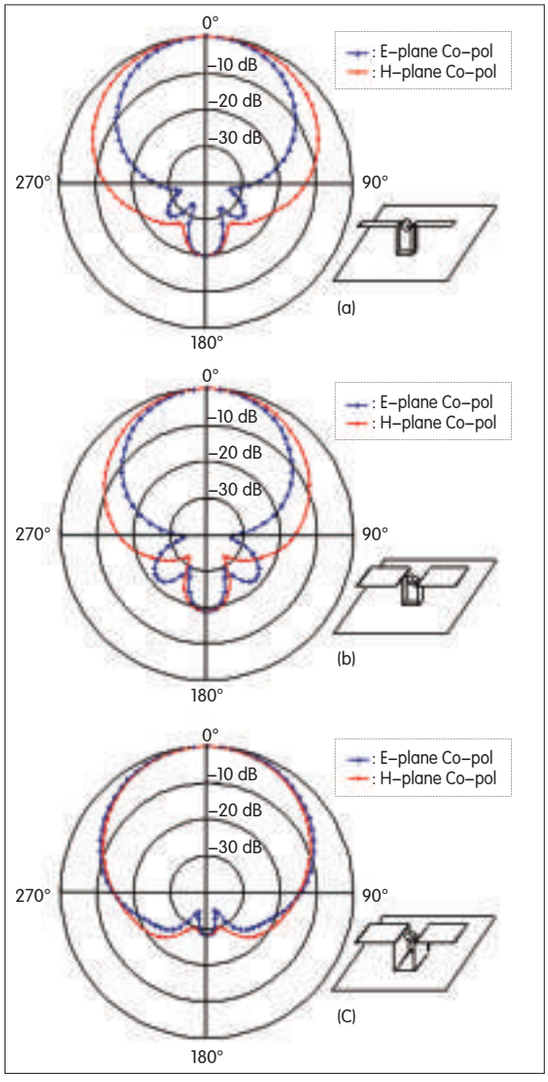
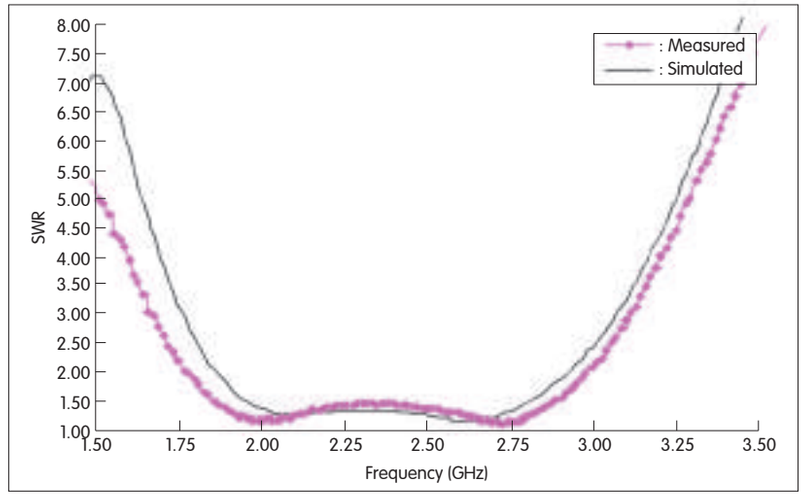


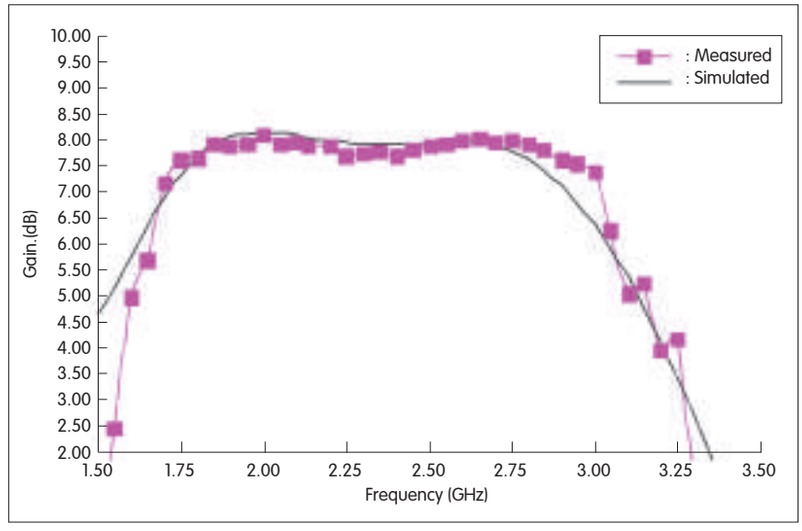
Figure 2.▶ A magneto–electric dipole.



▲Figure 3. Radiation patterns of (a) conventional dipole; (b) planar dipole; and (c) magneto–electric dipole.



▲Figure 4. Measured and simulated SWR against frequency for a wideband unidirectional antenna.



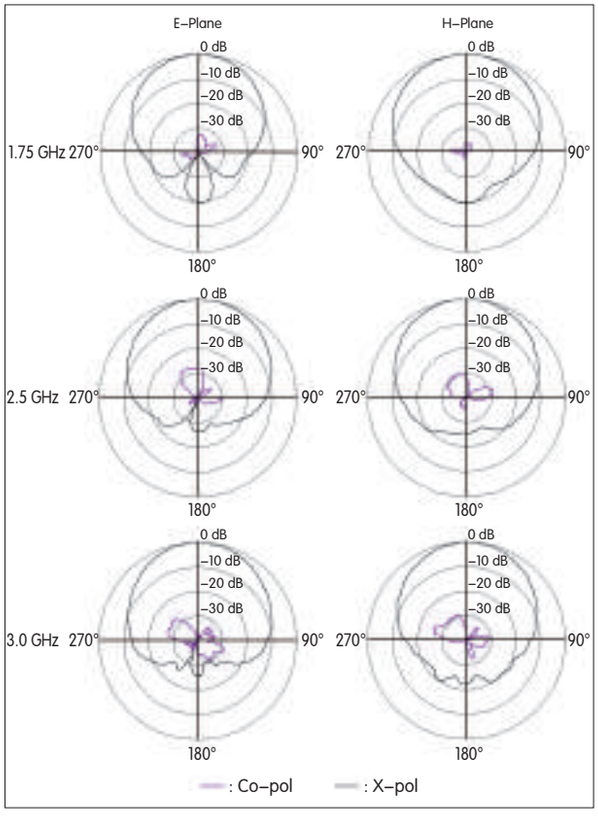
▲Figure 5. Measured and simulated gain against frequency for a wideband unidirectional antenna.

radiation and unequal E–plane and H–plane patterns. For the magneto–electric dipole in Fig.3c, the beamwidths in the E–plane and H–plane become almost identical. More importantly, the level of back radiation is significantly lower than that of the first and second cases by about 10 dB. The antenna can also maintain low cross–polarization.

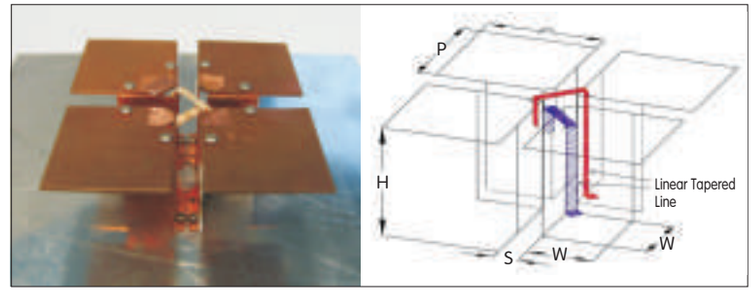
A prototype was built and tested. In Fig.4, the measured and simulated SWR curves show that the antenna has an impedance bandwidth of 52% (SWR≤2) when operating between 1.75 GHz and 3.0 GHz [41]. From the measured and simulated gain curves in Fig. 5, it can be seen that antenna gain varies between 7.5 dBi and 8.2 dBi across the operating frequency range, with an average value of 8 dBi [41]. Measured radiation patterns at frequencies of 1.75 GHz, 2.5 GHz, and 3 GHz are shown in Fig.6 [41]. The broadside radiation patterns in both E and H planes are stable and symmetric over the operating frequency range. At the center frequency of 2.5 GHz, the H–plane beamwidth is 79°, slightly larger than the E–plane beamwidth of about 75°. More importantly, there is low cross–polarization and low back radiation over the operating frequency range.

3 Dual–Polarized Magneto–Electric Dipole Antenna

Based on the linear polarized design, a dual–polarized magneto–electric dipole antenna was fabricated (Fig.7) [42]. This antenna also consists of electric dipoles and shorted patch antennas. The upper part of the antenna consists of two cross electric dipole elements with four square patches of dimensions $P \times P$=29.2 × 29.2 mm^2. The electric dipoles are



▲Figure 6. Measured radiation pattern at 1.75 GHz, 2.5 GHz, and 3.0 GHz.

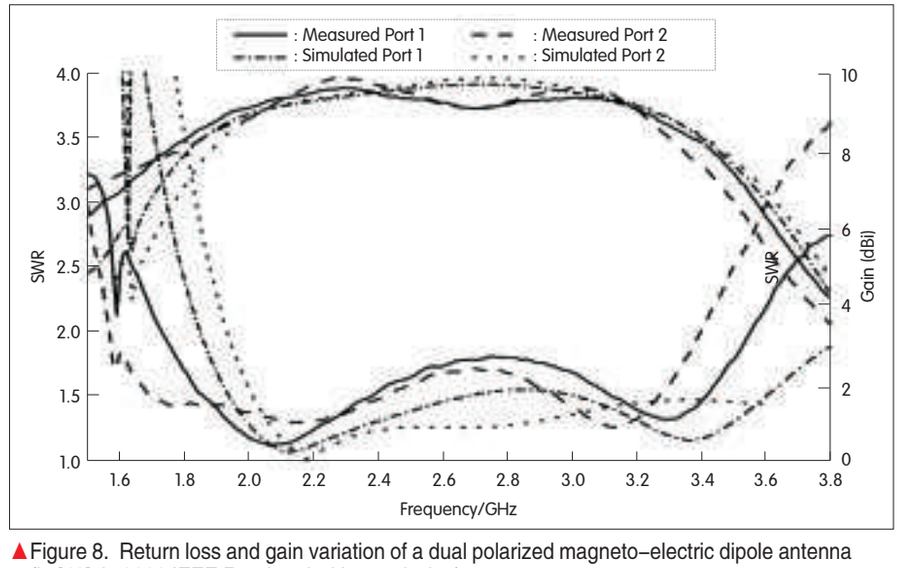


▲Figure 7. Geometry of dual–polarized magneto–electric dipole antenna (from [42] © 2009 IEEE Reprinted with permission).

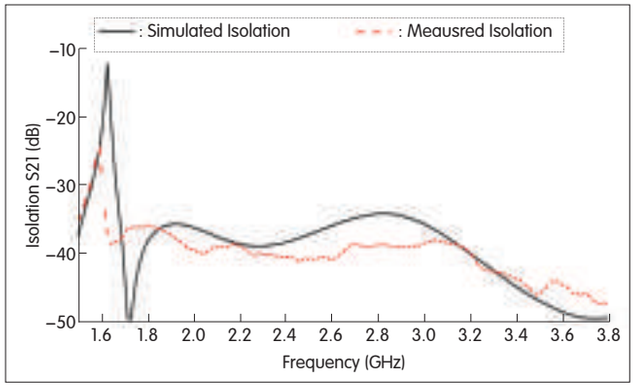
connected with the lower part of the antenna, which comprises four vertically oriented shorted patch antennas—with H=28 mm (~0.23 λ_0) and W=16.5 mm (~0.14 λ_0) — connected together. The distance between the two vertical walls of a shorted patch antenna is S=62 mm. Each Γ–shaped strip feed consists of a transmission line and a coupled strip. The transmission line, which is a linear tapered shape for impedance transformation, is located close to the corner of a folded vertical wall. The dimensions of the L–shaped coupling strip can be altered to achieve good impedance matching. The end of the coupling strip is not physically connected to the second arm of the dipole. It passes through the triangular hole of second arm of the dipole without connecting to the metal surface. The coaxial launcher underneath the ground plane is connected to the end of the linearly tapered line. To reduce mutual coupling between the two input ports, the two orthogonal coupling strips have different heights.

The prototype was measured with an HP8753ES network analyzer and a compact range measurement system. The results are shown in Fig. 8 and Fig. 9 [42]. The impedance bandwidths (SWR≤2) measured at ports 1 and 2 are 69.7% and 74.6% respectively. Because of the small difference in the dimensions of the coupling strips, the operating frequency ranges of the two polarizations are slightly different. As a result, the common bandwidth of the two ports is 65.9%, ranging from 1.7185 GHz to 3.409 GHz. Over the operating frequency range, the isolation between the two input ports is more than 36 dB. This fulfills the design requirement of mobile communications.

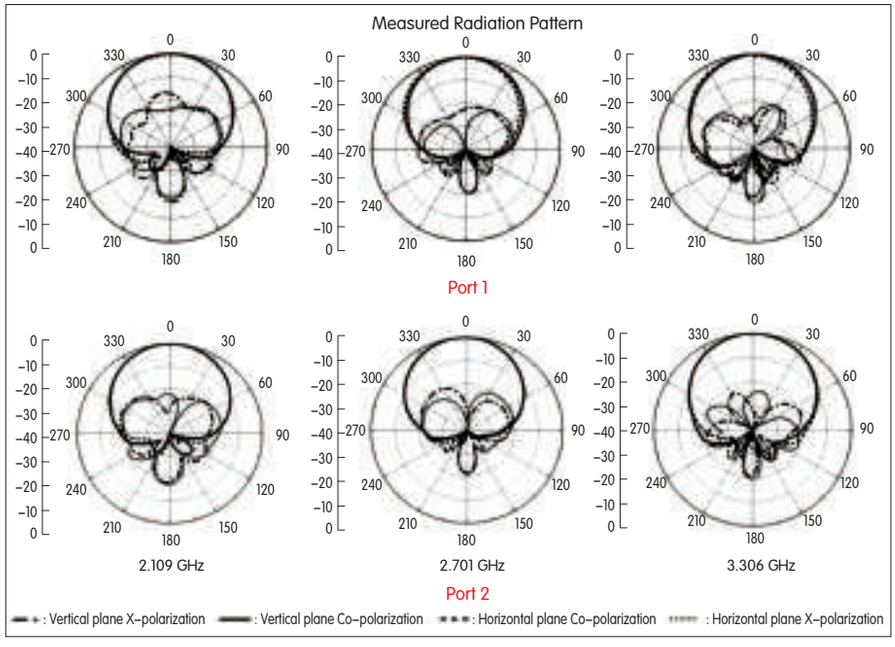
The variation in antenna gain is also shown in Fig.8. The gain is more than 9 dBi over a wide frequency range, with a maximum value of 9.5 dBi. More precisely, the 3 dB gain bandwidth covers a range from 1.6 to 3.55 GHz. The antenna's radiation patterns were obtained at frequencies of 2.109 GHz, 2.707 GHz, and 3.306 GHz (Fig.10) [42]. The radiation pattern is stable over the operating frequency range, and the back radiation is insignificant. The beamwidth varies only within a few degrees over the operating frequency range. If the size of the ground plane is reduced, the broadside radiation pattern does not change substantially, but back radiation and gain are affected. The antenna's performance is



▲Figure 8. Return loss and gain variation of a dual polarized magneto–electric dipole antenna (in [42] © 2009 IEEE Reprinted with permission).



◀Figure 9. Isolation variation of a dual polarized magneto-electric dipole antenna (in [42] © 2009 IEEE Reprinted with permission).



▲Figure 10. Radiation pattern of a dual polarized magneto-electric dipole antenna (in [42] © 2009 IEEE Reprinted with permission).

not very sensitive to changes in the dimensions of the gap and height of the antenna. This confirms the robustness of the design. Measurement results agree very well with simulation.

4 Horizontally Polarized Wideband Antenna with Conical Radiation Pattern

To create a horizontally polarized wideband antenna with conical radiation pattern, four magneto-electric dipoles are arranged in a ring, as shown in Fig.11 and Fig.12 [43]. The prototype is operated at 2 GHz. The dimensions shown in Table 1 were selected for excellent performance. The four magneto-electric dipoles are excited in-phase by a tapered power divider mounted underneath the ground plane. The upper part of the antenna is an electric dipole comprising a pair of sector-shaped horizontal plates. It is connected to the lower part of the antenna, which functions as a folded magnetic current. The magnetic current is provided by a vertically-oriented shorted patch antenna of height H=28 mm and that has a Γ-shaped strip feed. The separation between the two vertical plates is G =10.7286 mm. Each Γ-shaped strip has a transmission line section and a coupled strip section. The transmission line has a characteristic impedance of 50 Ω. The thickness of the transmission line is 0.3 mm, and the separation between the transmission line and nearby vertical plate is 1 mm.

The SWR and antenna gain are shown in Fig.13 [43]. The measured impedance bandwidth (SWR≤2) is about 38%, ranging from 1.61 GHz to 2.38 GHz. This is much larger than the bandwidth of many other designs of horizontally polarized conical beam antennas. The measured operating frequency band is slightly lower than the simulated operating frequency band by about 0.2 GHz, which is acceptable in practice. The measured gain is about 5 dBi on average, and the 3 dB gain bandwidth covers (and is larger than) the impedance bandwidth. The radiation patterns of the antenna at 1.6 GHz, 1.8 GHz, and 2 GHz are plotted in Fig.14 [43]. These are stable over the operating band. Compared with the simulation results, the measured radiation patterns have slightly higher cross-polarization and back lobe levels.

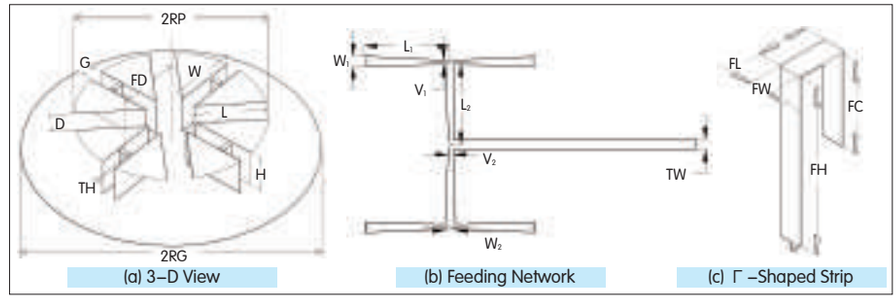
5 Conclusions

This paper begins by discussing the design of a magneto-electric dipole antenna. This structure has several advantages, including stable radiation pattern with low cross-polarization, low

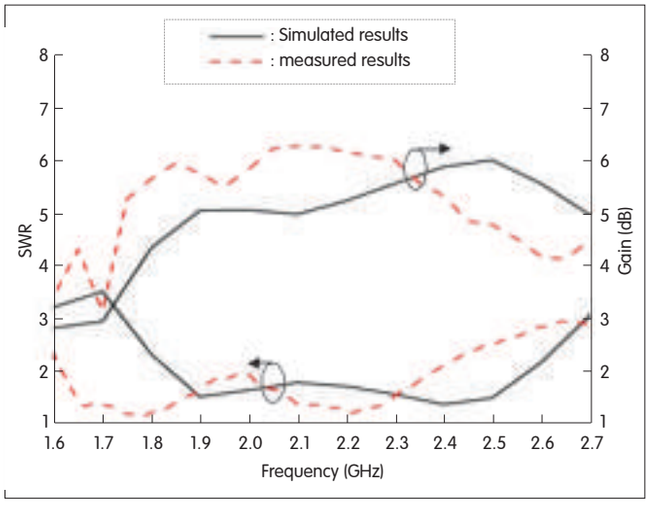
▼Table 1. Dimensions of a horizontally polarized wideband antenna

Parameters	D	L	TH	RP	RG
Values (mm)	17.3 (0.15 λ_o)	47.9 (0.32 λ_o)	1.0 (0.01 λ_o)	64.3 (0.43 λ_o)	100.0 (0.67 λ_o)
Parameters	W	G	FD	H	L1
Values (mm)	47.0 (0.31 λ_o)	10.7 (0.07 λ_o)	47.3 (0.32 λ_o)	28.0 (0.19 λ_o)	34.6 (0.23 λ_o)
Parameters	W1	V1	L2	W2	V2
Values (mm)	4.1 (0.03 λ_o)	1.5 (0.01 λ_o)	32.4 (0.22 λ_o)	4.1 (0.03 λ_o)	1.9 (0.01 λ_o)
Parameters	TW	FW	FL	FC	FH
Values (mm)	4.1 (0.03 λ_o)	3.5 (0.02 λ_o)	8.2 (0.06 λ_o)	14.7 (0.10 λ_o)	26.7 (0.18 λ_o)

▲Figure 11. Prototype of a horizontally polarized wideband antenna (in [43] © 2009 IEEE Reprinted with permission).

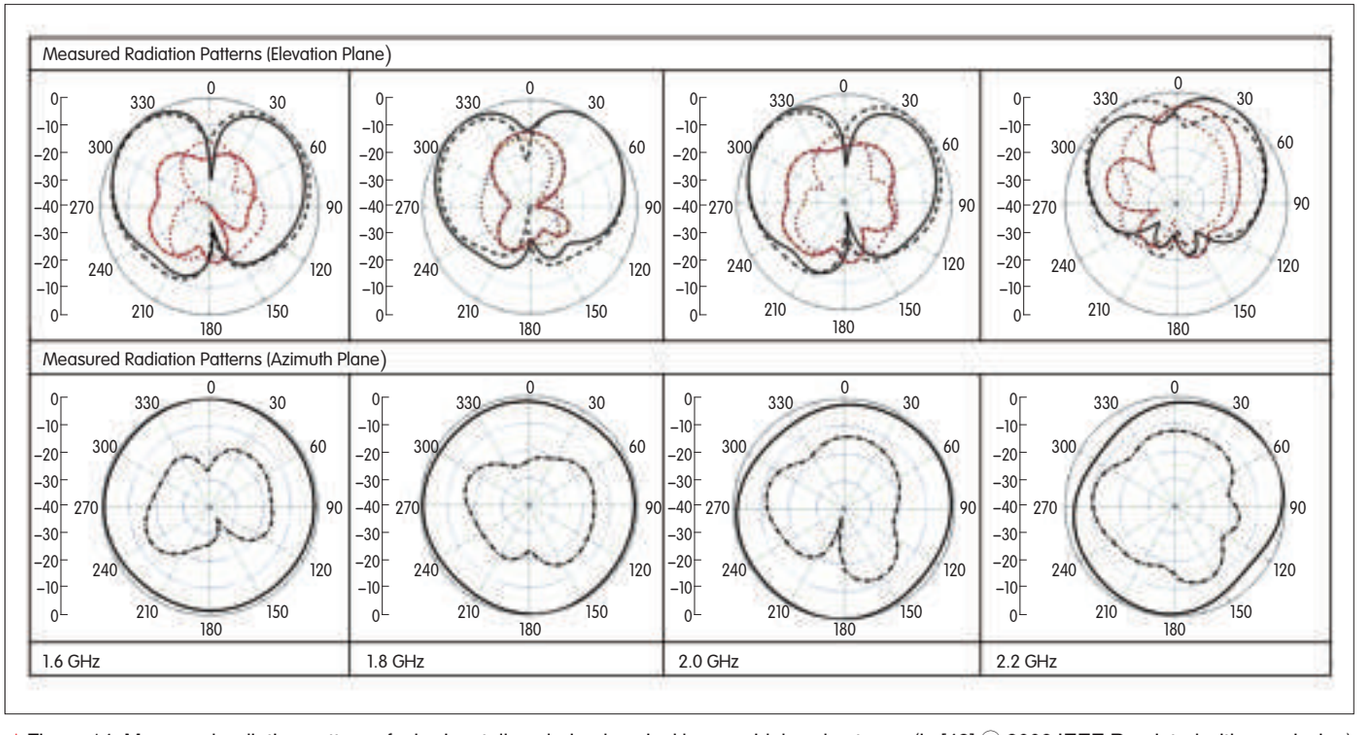


▲Figure 12. Geometry of a horizontally polarized wideband antenna (in [43] © 2009 IEEE Reprinted with permission).



◀Figure 13. SWR and antenna gain of a horizontally polarized conical beam wideband antenna (in Figure 14. Measured radiation pattern of a horizontally polarized conical beam wideband antenna (in [43] © 2009 IEEE Reprinted with permission).

radiation, nearly identical E-and H-plane patterns, and stable antenna gain over the entire operating frequency range. The proposed antenna has more than 52% impedance bandwidth (for SWR≤2) with a stable gain of 8dBi and low back radiation. A dual-polarized antenna element based on the magneto-electric dipole is also presented. This antenna has about 67% (for SWR≤2) impedance bandwidth, and its isolation is more than 36 dB over the impedance bandwidth. The maximum gain is 9.5 dBi. Furthermore, a horizontally polarized conical beam wideband



▲Figure 14. Measured radiation pattern of a horizontally polarized conical beam wideband antenna (in [43] © 2009 IEEE Reprinted with permission).

antenna is discussed. This antenna is low in profile (H=0.1867 λ_0) and has about 38% impedance bandwidth (for SWR≤2) and an average antenna gain of 5 dBi.

References

[1] A.M. Boifot, "Shortened, directive dipole for array antennas," *Int. J. Electron.*, 1991, 71, (1), pp. 127–137.
[2] C. M. Su, H. T. Chen, and K.L. Wong, "Printed dual–band dipole antenna with U–slotted arms for 2.4/5.2 GHz WLAN operation," *Electron. Lett.*, 2002, 38, pp. 1308–1309.
[3] E. Levine, S. Shtrikman and D. Treves, "Double–sided printed arrays with large bandwidth," *IEEE Proc. Microw.*, Antennas, Propag–Part H, vol.135, no.1, pp.54–59, 1988.
[4] S. Dey, P. Venugopalan, K. A. Jose, C.K. Aanandan, P. Mohanan and K.G. Nair, "Bandwidth enhancement by flared microstrip dipole antenna," in *IEEE Antennas and Propag. Soc. Int. Symp.* (AP–S), vol.1, London, Ontario, 1991, pp. 342–345.
[5] Y. D. Lin and S. N. Tsai, "Coplanar waveguide–fed uniplanar bow–tie antenna," *IEEE Trans. Antennas Propag.*, vol. 45, no. 2, pp.305–306, 1997.
[6] K. Kiminami, A. Hirata and T. Shiozawa, "Double–sided printed bow–tie antenna for UWB communications," *IEEE Antennas Wireless Propag. Lett.*, vol. 3, no. 1, pp.152–153, 2004.
[7] J. I. Kim, B. M. Lee and Y. J. Yoon, "Wideband printed dipole antenna for multiple wireless services," in *IEEE Radio and Wireless Conf.* (RAWCON'01), Boston, MA., 2001, pp.153–156.
[8] G. A. Evtioushkine, J. W. Kim and K. S. Han, "Very wideband printed dipole antenna array", *Electron. Lett.*, vol. 34, no. 2, pp. 2292–2293, 1998.
[9] K. M. Luk, C. L. Mak Y. Chow and K. F. Lee, "Broadband microstrip patch antenna," *Electron. Lett.*, vol. 34, no. 15, pp.1442–1443, 1998.
[10] C. L. Mak, K. M. Luk, K. F. Lee, and Y. L. Chow, "Experimental study of a microstrip patch antenna with an L–shaped probe," *IEEE Trans. Antennas and Propag.*, vol. 48, no. 5, pp. 777–783, 2000.
[11] Y. X. Guo, C. L. Mak, K. M. Luk, and K. F. Lee, "Analysis and design of L–probe proximity fed–patch antennas," *IEEE Trans. Antennas and Propag.*, vol. 49, no. 2, pp. 145–149, 2001.
[12] H. Wong, K. L. Lau and K. M. Luk, "Design of dual–polarized L–probe patch antenna arrays with high isolation," *IEEE Trans. Antennas Propag.*, vol. 52, no. 1, pp.45–52, 2004.
[13] H. Wong, and K. M. Luk, "A low–cost L–probe patch antenna array," in *Antennas Propag. Soc. Int. Symp.*, vol. 2, Boston, M.A., 2001, pp.280–282, 2001.
[14] F. Croq, and D. M. Pozar, "Millimeter wave design of wide–band aperture–coupled stacked microstrip antennas," *IEEE Trans. Antennas Propag.*, vol. 39, no.12, pp. 1770–1776, 1991.
[15] V. Rathi, G. kumar, and K. P. Ray, "Improved coupling for aperture coulped microstrip antennas," *IEEE Trans. Antennas Propag.*, vol. 44, No. 8, pp. 1196–1198, 1996.
[16] P. L. Sullivan, and D. H. Schaubert, "Analysis of an aperture coupled microstrip antenna," *IEEE Trans. Antennas Propag.*, vol. 34, no. 8, pp. 977–984, 1986.
[17] R. B. Waterhouse, "Design of probe–fed stacked patches," *IEEE Trans. Antennas Propag.*, vol. 47, no. 11, pp. 1780–1784, 1999.
[18] R. Q. Lee and K. F. Lee, "Experimental study of the two–layer electromagnetically coupled rectangular patch antenna," *IEEE Trans. AntennasPropag.*, vol. 38, no. 8, pp. 1298–1302, 1990.
[19] T. M. Au, and K. M. Luk, "Effect of parasitic element on the characteristics of microstrip antennas," *IEEE Trans. AntennasPropag.*, vol. 39, No. 8, pp. 1247–1251, 1991.
[20] H. Legay and L. Shafai, "New stacked microstrip antenna with large bandwidth and high gain," *IEEE Proc. Microw.*, AntennasPropag., vol. 141, no. 3, pp. 199–204, 1994.
[21] R. B. Waterhouse, "Design and scan performance of large, probe–fed stacked microstrip patch array," *IEEE Trans. Antennas Propag.*, vol. 50, no. 6, pp. 893–895, 2002.
[22] K. F. Lee, K. M. Luk, K. F. Tong, S. M. Shum, T. Huynh and R. Q. Lee, "Experimental and simulation studies of the coaxially fed U–slot rectangular patch antenna, " *IEEE Proc. Microw. Antenna Propag.*, vol. 144, no.5, pp. 354–358, 1997.
[23] T. Huynh, and K. F. Lee, "Single–layer single–patch wideband microstrip antenna," *Electron. Lett.*, vol. 31, no. 16, pp. 1310–1312, 1995.
[24] K. F. Lee, K. M. Luk, K. F. Tong, Y. L. Yung, and T. Huynh, "Experimental study of a two–element array of U–slot patches," *Electron. Lett.*, vol. 32, no. 5, pp. 418–420, 1996.
[25] M. Clenet, and L Shafai, "Multiple resonances and polarisation of U–slot patch antenna," *Electron. Lett.*, vol. 35, no. 2, pp. 101–103, 1999.
[26] K. F. Tong, K. M. Luk, K. F. Lee, and R. Q. Lee, "A broad–band U–slot rectangular patch antenna on a microwave substrate," *IEEE Trans. Antennas Propag.*, vol. 48, no. 6, pp. 954–960, 2000.
[27] Y. X. Guo, K. M. Luk, K. F. Lee, and Y. L. Chow, "Double U–slot rectangular patch antenna," *Electron. Lett.*, vol. 34, no. 19, pp. 1805–1806, 1998.
[28] A. Petosa, A. Ittipiboon and N. Gagnon, "Suppression of unwanted probe radiation in wideband probe–fed microstrip patches," *Electron. Lett.*, vol. 35, no. 5, pp. 355–357, 1999
[29] C. L. Mak, H. Wong and K. M. Luk, "High–gain and wide–band single–layer patch antenna for wireless communications, " *IEEE Trans. on Vehicular Tech.*, vol. 54, no. 1, 2005.
[30] H. W. Lai and K. M. Luk, "Design and study of wide–band patch antenna fed by meandering probe," *IEEE Trans. Antennas Propag.*, vol. 54, no. 2, 2006.
[31] H. W. Lai, and K. M. Luk, "Wideband patch antenna with low cross–polarisation," *Electron. Lett.*, vol. 40, no. 3, pp. 159–160, 2004.
[32] P. Li, H. W. Lai, K. M. Luk and K. L. Lau, "A Wideband patch antenna with cross–polarization suppression," *IEEE Antennas Wireless Propag. Lett.*, vol. 3, pp. 211– 214, 2004.
[33] H. W. Lai and K. M. Luk, "Wideband stacked patch antenna fed by a meandering probe," *Electron. Lett.*, vol. 41, pp. 297–298, 2005.
[34] H. W. Lai and K. M. Luk, "Wideband patch antenna fed by a modified L–shaped probe," *Microw. Optical Tech. Lett.*, vol. 48, no. 5, pp. 977–979, 2006.
[35] A. Clavin, "A new antenna feed having equal E– and H–plane patterns," *IEEE Trans. Antennas Propag.*, vol. 2, pp.113–119, 1954.
[36] A. Clavin, D. A. Huebner, and F. J. Kilburg, "An improved element for use in array antennas," *IEEE Trans. Antennas Propag.*, vol. 22, no.4, pp.521–526. 1974.
[37] R. W. P. King and G. H. Owyang, "The slot antenna with coupled dipoles, " *IRE Trans. Antennas Propag.*, vol. 8, pp.136–143, 1960.
[38] W. W. Black and A. Clavin, "Dipole augmented slot radiating element," U.S. Patent 3594806, Jul. 1971.
[39] W. F. Gabriel and L. R. Dod, "A complementary slot–dipole antenna for hemispherical coverage," NASA–Goddard Space Flight Center, Greenbelt, Md., NASA TM X–55681, Oct. 1966
[40] E. J. Wilkinson, "A circularly polarized slot antenna," *Microwave J.*, vol. 4, pp–97–100, Mar. 1961.
[41] Kwai Man Luk and Hang Wong, "A New Wideband Unidirectional Antenna Element," *Int. J. Microw. Optical Tech.*, vol. 1, no. 1, pp. 35– 44, 2006.
[42] Bi Qun Wu, and Kwai–Man Luk, "A Broadband Dual–Polarized Magneto–Electric Dipole Antenna With Simple Feeds," *IEEE Antennas Wireless Propag. Lett.*, vol. 8, pp. 60–63, 2009.
[43] Bi Qun Wu, and Kwai–Man Luk, "A Wideband, Low–Profile, Conical–Beam Antenna With Horizontal Polarization for Indoor Wireless Communications," *IEEE Antennas and Wireless Propag. Lett.*, vol. 8, pp. 634–636, 2009.

Biographies

Hang Wong (hang.wong@cityu.edu.hk) received his B.Eng., M.Phil., and Ph.D. degrees in electronic engineering from the City University of Hong Kong. He joined the Wireless Communications Research Center (RCW) at City University of Hong Kong in 2002 as an antenna engineer. He is currently a senior engineer at the State Key Laboratory (SKL) of Millimeter Waves, in Hong Kong. His research interests include design of broadband antennas, RFID antennas, small antennas, GPS antennas, millimeter wave antennas, and antenna arrays. He is the author of chapters in books on antenna research. He was the co–inventor of linear/circularly–polarized, dual–polarized, and small printed antennas and has been awarded patents on these in the U.S. and PRC. Dr. Wong was awarded the Outstanding Research Thesis Award from City University of Hong Kong in 2002. He received the Microwave Student Prize at the Asia Pacific Microwave Conference 2006 held in Yokohama and received the Best Paper Award at the International Symposium on Antennas and Propagation 2008 in Taipei.

Kwai–Man Luk (eekmluk@cityu.edu.hk) received his B.Sc.(Eng.) and Ph.D. degrees in electrical engineering from the University of Hong Kong. He joined the Department of Electronic Engineering at City University Hong Kong in 1985 as a lecturer. Two years later, he moved to the Department of Electronic Engineering at the Chinese University of Hong Kong where he spent four years. Professor Luk returned to the City University Hong Kong in 1992, and he is currently chair professor of Electronic Engineering and director of the State Key Laboratory in Millimeter waves (Hong Kong). His research interests include design of patch, planar and dielectric resonator antennas, and microwave measurements. He is the author of three books, nine book chapters, more than 260 journal papers, and 200 conference papers. He has been awarded two U.S. patents and more than 10 PRC patents on the design of a wideband patch antenna with L–shaped probe feed. He was the technical program chairperson of the 1997 Progress in Electromagnetics Research Symposium (PIERS 1997), the general vice–chairperson of the 1997 and 2008 Asia–Pacific Microwave Conference, and the general chairman of the 2006 IEEE Region Ten Conference. Professor Luk received the Japan Microwave Prize at the 1994 Asia Pacific Microwave Conference held in Chiba in December 1994 and the Best Paper Award at the 2008 International Symposium on Antennas and Propagation held in Taipei in October 2008. He was awarded the 2000 Croucher Foundation Senior Research Fellow in Hong Kong. He is a deputy editor–in–chief of JEMWA. Professor Luk is a Fellow of the Chinese Institute of Electronics, PRC, a Fellow of the Institution of Engineering and Technology, UK, a Fellow of the Institute of Electrical and Electronics Engineers, USA, and a Fellow of the Electromagnetics Academy, USA.

Advanced Synthesis Techniques for Microwave Filters

Richard J Cameron
(Canopus Consultancy)

Abstract: With the advent of the 'digital revolution' that has made possible services such as the world wide web, satellite broadcasting and mobile and trunk telephony, the finite RF spectrum allocated for terrestrial and satellite telecommunication systems is becoming increasingly crowded. This has impacted significantly upon the performance required from the microwave equipment that comprises these systems. In the case of microwave filters, greater in–band linearity to avoid signal distortion and out–of–band isolation to suppress interference are routinely specified, which can only be satisfied by advanced filtering characteristics. This article presents the coupling matrix approach to the synthesis of prototype filter networks, enabling the realization of the hardware embodying the enhanced performance needed by today's high capacity systems.

Keywords: filter network synthesis; coupling matrix; microwave filters

1 Introduction

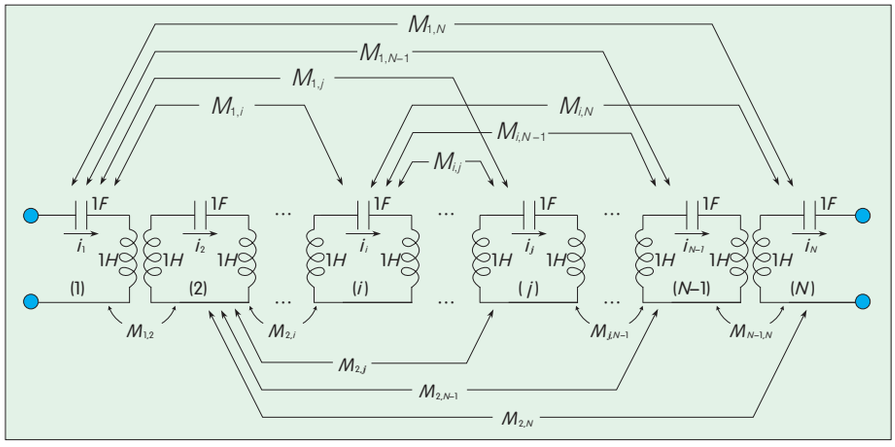
Until the early 1970s, nearly all filter synthesis techniques were based on the extraction of electrical elements—lumped capacitors and inductors, and transmission line lengths—from the polynomials that represented the filter's electrical performance in mathematical terms. This was perfectly adequate for the technologies and applications that were available at the time. Many important contributions were made to the art of advanced filter transfer and reflection polynomial generation and to their conversion into electrical component values corresponding to the filter technologies that were available in those days [1]–[3].

In the early 1970s, the first satellite telecommunication systems were in operation, and demand for their services was growing enormously. This meant that RF spectrum allocated to satellite communication systems had to be pushed to higher frequency bands in order to accommodate the increasing volumes of traffic. The technology available to implement components of these higher–frequency systems was also advancing; for example, better front–end low–noise amplifiers, high power transmit amplifiers, antenna systems, and passive channelizing equipment. Crowding of the available spectrum meant that the specifications for channel filters in terms of in–band linearity (group delay, insertion loss) and out–of–band selectivity (high close–to–band rejection; and for transmit filters, lowest possible insertion loss) became more demanding.

During this period, two important advances were made in the field of filter design to address the new demands. The first was the development of design methods for advanced filtering functions incorporating built–in transmission zeros and group delay features aimed particularly at microwave filter implementation. Then, the 'reflex' (sometimes called 'folded') cross–coupled microwave filter [4] was introduced, which allowed inter–resonator couplings other than the usual main–line couplings between sequentially–numbered resonators to be implemented. These cross–couplings, as they came to be known, enabled the realization of special features of a filtering function, namely, transmission zeros to give a high close–to–band rejection of RF noise and interference, or linearization of in–band group delay, or both within the same filter structure.

The other major advance around this time was the development of dual–mode technology for waveguide filters at ComSat Laboratories [5], in response to very stringent performance requirements being imposed on spaceborne microwave equipment by system designers. The innovation came in two parts–firstly the development of the coupling matrix method for the holistic design of the filter's main and cross–coupling elements, and secondly the 'propagating' dual–mode waveguide configuration which inherently provided the cross–couplings necessary for the realization of the special performance features, without the need for complex and sensitive coupling elements.

Richard J Cameron



▲Figure 1. Multicoupled network—a classical "bandpass prototype" representation (courtesy A. E. Atia).

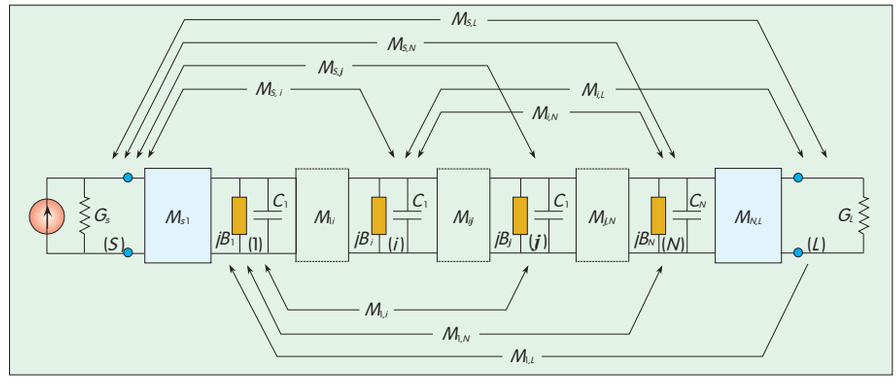
Since the 1970s, the coupling matrix has become the microwave filter design tool of choice—for the initial design and then for the tuning, modeling, and analysis microwave filter performance. One important feature is the one–to–one correspondence between individual physical components of the filter and the elements of the coupling matrix. Although the initial design of a filter network assumes frequency–independent coupling elements as well as lossless and dispersionless resonators, these real–world effects may be accommodated when analyzing the matrix for filter performance prediction Different characteristics may be allocated to different elements if there is a mix of technologies in the filter. Another advantage is the ability to reconfigure the coupling matrix through similarity transforms to arrive at a different coupling arrangement that corresponds to the available coupling elements of the particular microwave structure selected for the application. This can be done without going right back to the beginning of the network synthesis process and starting again on a different network synthesis route. This would be necessary if a classical element extraction method were used. Coupling matrix synthesis theory has been advanced to include asymmetric filtering characteristics, which have become important for terrestrial telecom systems, particularly mobile telephony systems.

Because of the prevalence of the coupling matrix in microwave filter design, this article will concentrate on techniques for the synthesis of and then the reconfiguration of the coupling matrix ready for realization in a variety of microwave structures. First, the method for the generation of advanced polynomial filtering functions will be briefly outlined followed by the synthesis of one of the canonical networks—the transversal matrix. Then, reconfiguration of the transversal matrix into various forms for realization in a variety of microwave structures will be discussed. Some examples are given to clarify aspects of the design processes, and references cited if further information is required by the reader.

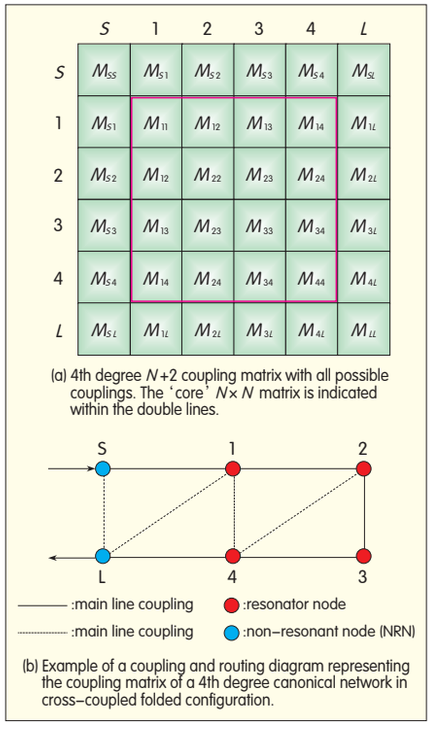
2 The Coupling Matrix

The basic circuit model that was used in [5] was a 'bandpass prototype,' which is a generalized multicoupled network as shown in Fig. 1. The circuit comprises a cascade of lumped element series resonators intercoupled through transformers. Each resonator comprises a 1F capacitor in series with the self inductances of the main–line transformers, which total 1 H within each loop. This gives a centre frequency of 1 rad/s, and the couplings are normalized to give a bandwidth of 1 rad/s. In addition, every loop is theoretically coupled to every other loop through cross–mutual couplings between the main–line transformers.

This network may be represented by an $N \times N$ coupling matrix where N is the number of resonators (the degree or order of the filter). The elements of the matrix contain the values of the couplings between each of the resonators; between sequentially–numbered resonator nodes (main–line couplings), and non–adjacent nodes (cross–couplings). Because the electrical elements of the network are passive and reciprocal, the matrix is symmetrical about its principal diagonal. To more closely represent a microwave circuit, the transformers may be replaced by immittance inverters (90° lengths of transmission line), which approximates the electrical characteristic of many microwave coupling devices. By placing an inverter at each end of the network, the input and output couplings of the filter may also be represented (Fig. 2). With



▲Figure 2. Multicoupled network—equivalent lowpass prototype modified to include FIRs and immittance inverters.



(a) 4th degree N+2 coupling matrix with all possible couplings. The 'core' $N\times N$ matrix is indicated within the double lines.

(b) Example of a coupling and routing diagram representing the coupling matrix of a 4th degree canonical network in cross–coupled folded configuration.

▲Figure 3. N+2 coupling matrix.

the extra inverters, the matrix increases to (N+2) × (N+2) in size—the so–called 'N+2' coupling matrix—and becomes the dual network in Fig. 1.

This circuit as it stands only supports symmetric filtering characteristics. But with the addition of a series–connected frequency–invariant reactance (FIR) within each loop, the capability of the circuit may be extended to include asymmetric cases (Fig. 2). These have been finding increasing application recently as the RF frequency spectrum becomes more crowded and rejection specifications more severe.

The FIR—sometimes referred to as a 'self' coupling—represents a frequency offset of the resonator it is associated with, and its value is entered along the diagonal of the coupling matrix. Because the inverters are also frequency–invariant and there are no self–inductors, the network in Fig. 2 may now be considered as a lowpass prototype, which simplifies the synthesis process somewhat.

The N+2 short–circuit admittance matrix[y']for the network in Fig. 2 may be separated out into its purely resistive and purely reactive parts:

$$[y']=[\mathrm{G}]+[j\mathbf{M}+\mathbf{U}]=[\mathbf{G}]+[y] \quad (1)$$

where the purely real matrix [**G**] contains the conductive terminations G_S and G_L of the network and the purely reactive admittance $[y]=[j\mathbf{M}+\mathbf{U}]$ is the sum of the coupling matrix **M** and the diagonal matrix **U** which contains the frequency variable $s\,(=j\omega)$, except for U_{SS} and U_{LL} which are zero.

The N+2 coupling matrix [**M**] contains the values of all the couplings in the network, including the input/output couplings (which may connect to internal resonators). The diagonal contains the values of the frequency invariant reactances that represent resonator frequency offsets (the negative values of FIRs in Fig. 2), which are necessary for asymmetric characteristics. Fig. 3(a) shows a canonical 4th degree coupling matrix with all couplings present. Fig. 3(b) is an example of a typical coupling and routing diagram, representing a possible inter–resonator coupling arrangement for the 'folded' topology.

3 Synthesis Procedure

The filter design process begins with the generation of the rational polynomials embodying the transfer and reflection characteristics $S21$ and $S11$ that satisfy the rejection and in–band specifications of the application. Once the polynomials have been obtained, the next step in the synthesis process is to synthesize the coupling matrix and configure it so that its non–zero entries coincide with the available coupling elements of the structure it intends to use for realizing the filter response. Finally, the dimensions of the coupling elements are calculated from the coupling matrix values.

The procedure is illustrated in Fig. 4 for a 6th degree characteristic with two transmission zeros and realized in coupled waveguide resonator technology. The direct correspondence

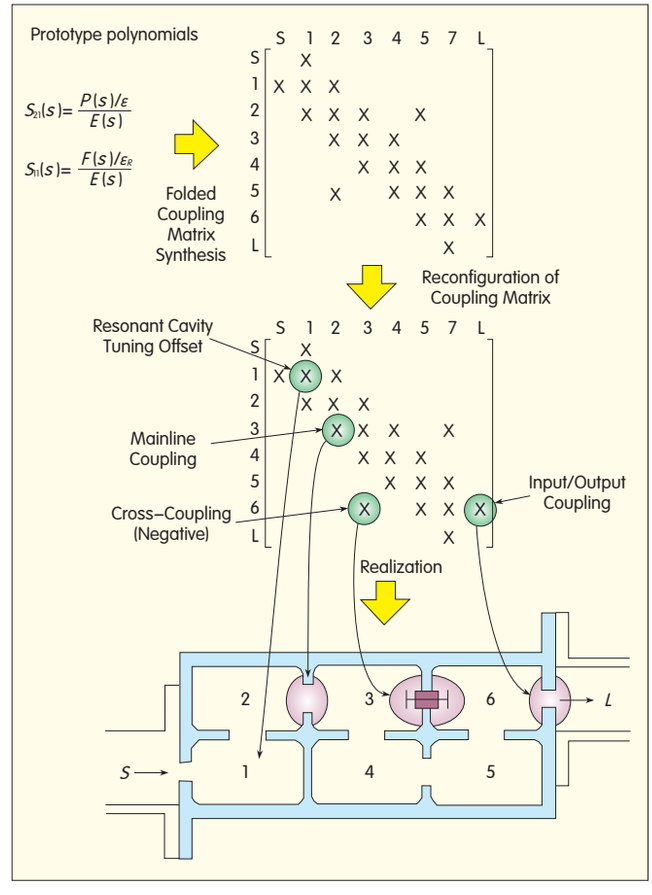
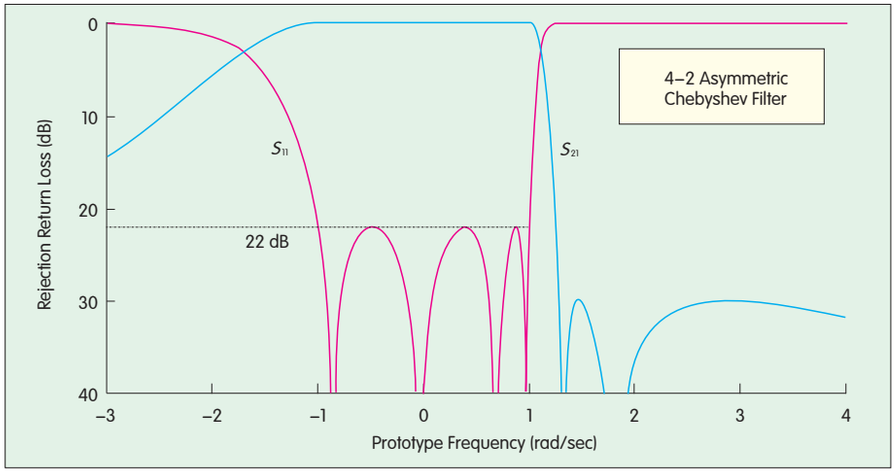


Figure 4. ▶ Microwave filter design process: synthesis of the polynomials for the transfer and reflection function, synthesis of canonical coupling matrix, reconfiguration of coupling matrix, realization in microwave coupled–resonator technology.

Richard J Cameron



▲Figure 5. Lowpass prototype transfer and reflection characteristics of the 4–2 asymmetric Chebyshev filter with two prescribed transmission zeros at $s_{01} = +j1.3217$ and $s_{02} = +j1.8082$.

between the elements of the coupling matrix and the physical filter components is indicated.

3.1 Generation of Transfer and Reflection Polynomials

In modern telecommunication, radar, and broadcast systems, where the allocated RF frequency spectrum has become very congested, the specifications on performance from the component microwave filters have become increasingly stringent. For these applications, Chebyshev class of filtering characteristic is very suitable on account of the inherent equiripple in–band return loss level and the ability to build in transmission zeros (TZs) to provide high close–to–band rejection levels, or in–band group delay equalization, or both within the same filtering function. Moreover, the TZs may be placed asymmetrically to optimally comply with asymmetric specifications. A method for generating the lowpass prototype polynomials for the Chebyshev class filter function is outlined below.

For any two–port lossless filter network composed of a series of N intercoupled resonators, the transfer and reflection functions may be expressed as a ratio of two polynomials [6]:

$$S_{11}(\omega)=\frac{F(\omega)/\varepsilon_R}{E(\omega)} \quad S_{21}(\omega)=\frac{P(\omega)/\varepsilon}{E(\omega)} \qquad (2)$$

$$\text{where } \varepsilon=\frac{1}{\sqrt{10^{RL/10}-1}}\cdot\left|\frac{P(\omega)}{F(\omega)/\varepsilon_R}\right|_{\omega=\pm 1}$$

and RL is the prescribed inband equiripple return loss level of the Chebyshev function in dB. $S_{11}(\omega)$ and $S_{21}(\omega)$ share a common denominator $E(\omega)$ The polynomials $E(\omega)$ and $F(\omega)$ are both of degree N, when the polynomial $P(\omega)$ carries the n_{fz} transfer function finite–position transmission zeros. For a Chebyshev filtering function, ε is a constant normalizing $S_{21}(\omega)$ to the equiripple level at $\omega=\pm 1$, and ($\varepsilon_R = 1$ except for fully canonical filters (ie. $n_{fz} = N$).

For a prescribed set of transmission zeros that make up the polynomial $P(\omega)$ and a given equiripple return loss level, the reflection numerator polynomial $F(\omega)$ may be built up with an efficient recursive technique. and then the polynomial $E(\omega)$ found from the conservation of energy principle [6].

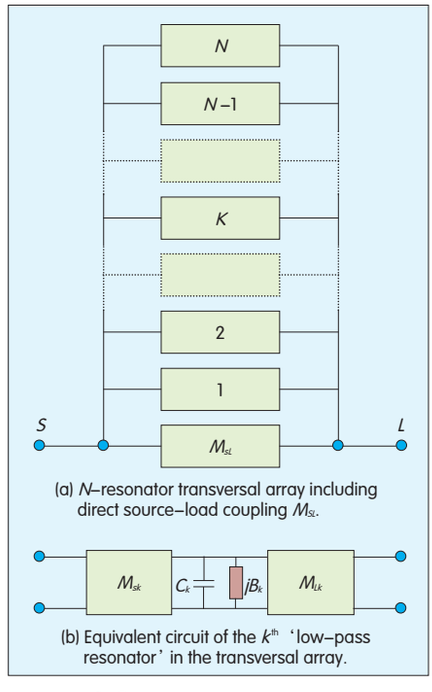
An example of this synthesis method is given in [6] for a 4th degree prototype with 22 dB return loss level and two imaginary axis TZs at $s_{01} = +j1.3127$ and $s_{02} = +j1.8082$. These are positioned to give two rejection lobes at 30 dB each on the upper side of the passband. Plots of the transfer and rejection characteristics are shown in Fig. 5.

3.2 Construction of the $N+2$ Transversal Matrix

The second step in the synthesis procedure is to calculate the values of the coupling elements of a canonical coupling matrix from the transfer and reflection polynomials. Three forms of the canonical matrix are commonly used–the folded [4], transversal [7] or arrow [8]. The transversal matrix is particularly easy to synthesize, and the other two may be derived from it quite simply by applying a formal series of analytically–calculated similarity transforms.

The transversal coupling matrix comprises a series of N individual 1st degree low pass sections, connected in parallel between the source and load terminations but not to each other (Fig. 6 (a)). The direct source–load coupling inverter M_{SL} is included to allow fully canonical transfer functions to be realized according to the "minimum path" rule, i.e. n_{fzmax}, the maximum number of finite–position TZs that may be realized by the network=$N-n_{min}$, where nmin is the number of resonator nodes in the shortest route through the couplings of the network between the source and load terminations. In fully canonical networks, $n_{min} = 0$ and So $n_{fzmax} = N$ (the degree of the network).

Each N low–pass section comprises one parallel–connected capacitor C_k and one frequency invariant susceptance B_k, connected through admittance inverters of characteristic



(a) N–resonator transversal array including direct source–load coupling M_{SL}.

(b) Equivalent circuit of the k^{th} 'low–pass resonator' in the transversal array.

▲Figure 6. Canonical transversal array.

admittances M_{Sk} and M_{Lk} to the source and load terminations respectively. The circuit of the k^{th} lowpass section is shown in Fig. 6(b).

The approach employed to synthesize the N+2 transversal coupling matrix is to construct a 2–port short–circuit admittance parameter matrix $[\boldsymbol{Y}_N]$ for the overall network in two ways: from the coefficients of the rational polynomials of the transfer and reflection scattering parameters $S_{21}(s)$ and $S_{11}(s)$ (which represent the characteristics of the filter to be realized) or from the circuit elements of the transversal array network. By equating the $[\boldsymbol{Y}_N]$ matrices derived by these two methods, the elements of the coupling matrix associated with the transversal array network can be related to the coefficients of the $S_{21}(s)$ and $S_{11}(s)$ polynomials [7].

An example of a reciprocal N+2 transversal coupling matrix M representing the network is shown in Fig.7. M_{Sk} are the N input couplings, and they occupy the first row and column of the matrix from positions 1 to N. Similarly, M_{Lk} are the N output couplings, and they occupy the last row and column of M from positions 1 to N. All other entries are zero.

4 Similarity Transformation and Reconfiguration

The elements of the transversal coupling matrix that result from the synthesis procedure can be realized directly by the coupling elements of a filter structure if it is convenient to do so. However, for most coupled–resonator technologies, the couplings of the transversal matrix are physically impractical or impossible to realize. It becomes necessary to reconfigure the matrix with a sequence of similarity transforms (sometimes called rotations) [8] until a more convenient coupling topology is obtained. The use of similarity transforms ensures that the eigenvalues and eigenvectors of the matrix M are preserved. Under analysis, the transformed matrix yields exactly the same transfer and reflection characteristics as the original matrix.

There are several more practical canonical forms for the transformed coupling matrix M. Two of the better–known forms are the 'arrow' form [8] and the more generally useful 'folded' form [4]. Either of these canonical forms can be used directly if it is convenient to realize the couplings or be used as a starting point for the application of further transforms to create an alternative resonator intercoupling topology optimally adapted to the physical and electrical constraints of the technology with which the filter will eventually be realized. The method for reduction of the coupling matrix to the folded form with a formal sequence of rotations is detailed in [6]. The 'arrow' form may be derived using a very similar method.

5 Advanced Configurations

In this section, some advanced coupling matrix configurations particularly suitable for filters and diplexers in terrestrial telecommunication systems will be considered. An important application is in the cellular telephony industry where strong growth has meant that very stringent out–of–band rejection and in–band linearity specifications have had to be imposed to cope with a crowded frequency spectrum and increasing numbers of channels. At the RF frequencies allocated to mobile systems (L–band, S–band, and sometimes C–band), coaxial or dielectric resonator technology is often used for the filters of the system because of the compact, flexible, and robust construction with flexible layout possibilities that may be achieved together with the ability to realize advanced filtering characteristics and quite high RF power handling.

A microwave filter topology that has found widespread application in both terrestrial and space systems is the 'trisection.' The basic trisection may be used as a stand–alone section or be embedded within a higher–degree filter network. But often multiple trisections are merged to form advanced configurations such as cascaded 'N–tuplets' or box filters.

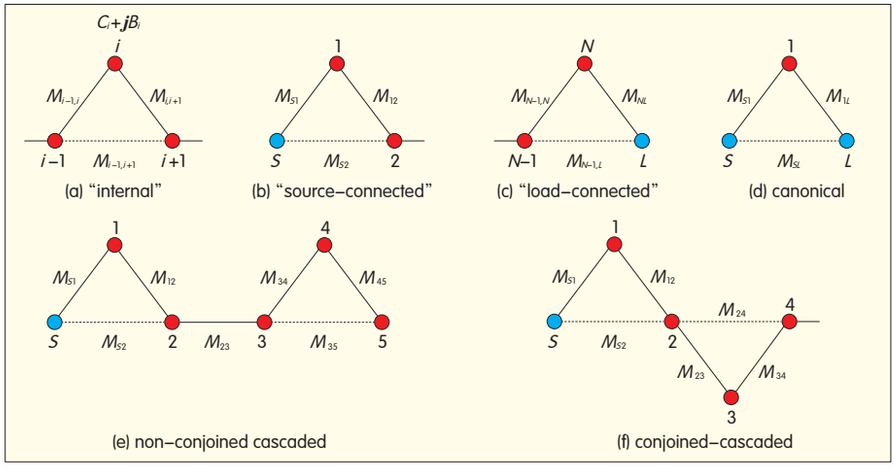
5.1 Trisections

A trisection comprises three couplings between three sequentially–numbered nodes of a network (the first and third of which may be source or load terminals) or it might be embedded within the coupling matrix of a higher–degree network [9]. The minimum path rule indicates that trisections are able to realize one transmission zero each. As will be shown later, trisections may be merged using rotations to form higher–order sections; for example, a quartet capable of realizing two TZs can be formed by merging two trisections.

Fig.8 shows four possible configurations. Fig.8(a) is an internal trisection, whilst Figs.8(b) and (c) show 'input' and 'output' trisections respectively, where one node is the source or load termination. When the

	S	1	2	3	⋯	k	⋯	N–1	N	L
S		M_{S1}	M_{S2}	M_{S3}	⋯	M_{Sk}	⋯	$M_{S,N-1}$	M_{SN}	M_{SL}
1	M_{1S}	M_{11}								M_{1L}
2	M_{2S}		M_{22}							M_{2L}
3	M_{3S}			M_{33}						M_{3L}
⋮	⋮				⋱					⋮
k	M_{kS}					M_{KK}				M_{KL}
⋮	⋮						⋱			⋮
N–1	$M_{N-1,S}$							$M_{N-1,N-1}$		$M_{N-1,L}$
N	M_{NS}								M_{NN}	M_{NL}
L	M_{LS}	M_{L1}	M_{L2}	M_{L3}	⋯	M_{LK}	⋯	$M_{L,N-1}$	M_{LN}	

▲Figure 7. N+2 Canonical coupling matrix M for the transversal array. The 'core' $N \times N$ matrix is indicated within the double lines. The matrix is symmetric about the principal diagonal i.e. $M_{ij}=M_{ji}$



▲Figure 8. Coupling and routing diagrams for trisections.

first and third nodes are the source and load terminations respectively (Fig. 8(d)), we have a canonical network of degree 1 with the direct source–load coupling, M_{SL} , providing the single transmission zero. Trisections may also be cascaded with other trisections, either separately or conjoined (Figs. 8(e) and (f)).

Beingable to realize just one transmission zero each, the trisection is very useful for synthesizing filters with asymmetric characteristics. They may exist singly within a network or multiply as a cascade. Rotations may be applied to reposition them along the diagonal of the overall coupling matrix or to merge them to create quartet sections (two trisections) or quintet sections (three trisections). The following is an efficient procedure for synthesizing a cascade of trisections [9].

5.2 Synthesis of the 'Arrow' Canonical Coupling Matrix

The folded cross–coupled circuit and its corresponding coupling matrix was previously introduced as one of the basic canonical forms of the coupling matrix. It is capable of realizing N transmission zeros in an Nth degree network. A second form was introduced by Bell [8] in 1982, which later become known as the 'wheel' or 'arrow' form. Like the folded form, all the main–line couplings are present; and in addition, the source terminal and each resonator node is cross–coupled to the load terminal.

Fig. 9(a) is an example of a coupling and routing diagram for a 5th degree canonical filtering circuit. It shows clearly why this configuration is referred to as the 'wheel.' with the main–line couplings forming the (partially incomplete) rim and the cross–couplings and input/output coupling forming the spokes. Fig. 9(b) shows the corresponding coupling matrix where the cross–coupling elements are all in the last row and column, and together with the main line and self couplings on the main diagonals give the matrix the appearance of an arrow pointing downwards towards the lower right corner of the matrix. The arrow matrix may be synthesized from the canonical transversal matrix with a formal sequence of rotations, similar to that of the folded matrix.

The basis of the trisection synthesis procedure relies on the fact that the value the determinant of the self and mutual couplings of the trisection evaluated at $\omega=\omega_0$ (the position of the TZ associated with the trisection) is zero:

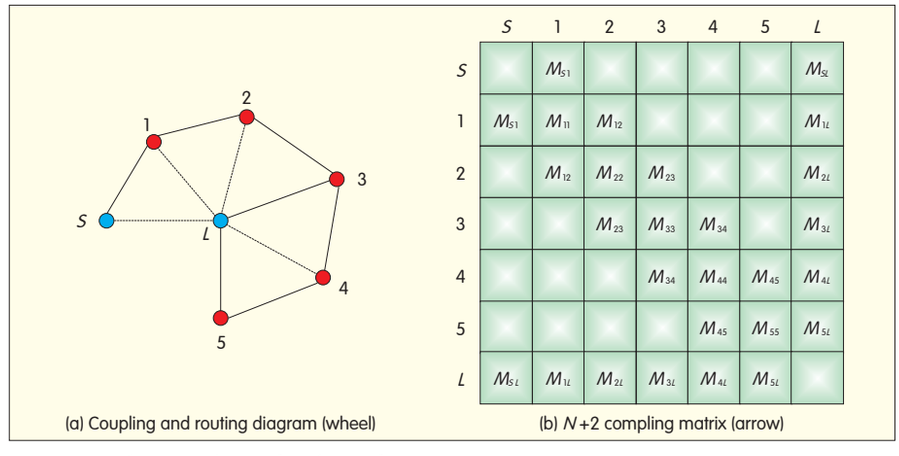
$$\det\begin{vmatrix} M_{k-1,k} & M_{k-1,k+1} \\ \omega_0+M_{k,k} & M_{k,k+1} \end{vmatrix}=0 \qquad (3)$$

where k is the number of the middle resonator of the trisection. Knowing the positions of the transmission zeros of the filtering characteristic, the trisections can be generated one by one within the arrow matrix, and shifted to form a cascade between the input and output nodes.

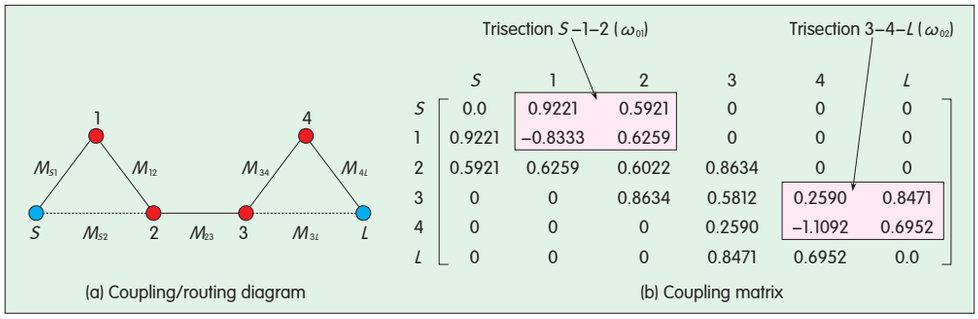
Fig. 10 gives the topology and coupling matrix for the 4th degree filter with 22 dB RL and two transmission zeros at (ω_{01}=1.8082 and (ω_{02}=1.3217 that was used as an example above now configured with two trisections (to realize the two TZs). The shaded areas in the matrix indicate the couplings associated with each trisection.

Once the arrow coupling matrix has been formed, the procedure to create the first trisection realizing the first TZ at $\omega=\omega_{01}$ begins with conditioning the matrix with the application of a rotation at pivot [N–1, N] and an angle (θ_{01} to the original arrow matrix $M^{(0)}$. This trisection is then shifted by a series of rotations to the left of the network.

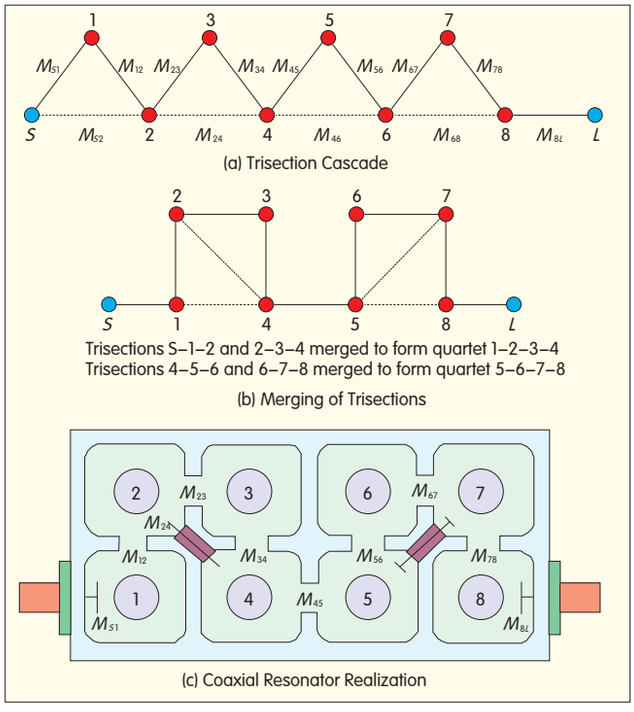
Now he process can be repeated for the second trisection at $\omega=\omega_{02}$ and so



▲Figure 9. 5th degree 'wheel' or 'arrow' canonical circuit.



▲Figure 10. 4th degree filter with two transmission zeros realized as trisections.



◀Figure 11. 8–4 asymmetric filter.

on until a cascade of trisections is formed—one for each of the TZs in the original prototype, as shown in Fig. 11(a). The trisections may be realized directly if it is convenient to do so; for example, for coupled coaxial resonators. But for other technologies such as dual–mode waveguide, a cascade of quartets may be more suitable. A cascade of quartets is easily achieved by merging adjacent trisections, as illustrated in Fig. 11(b). Fig.11(c) shows a possible coaxial–resonator realization for the two quartets.

This procedure can be extended to form even higher–order sections in cascade; for example, three trisections may be merged to form a quintet section, as illustrated in Fig.12.

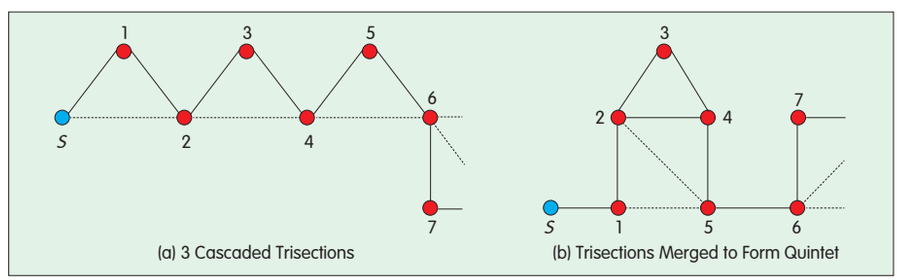
6 Box and Extended Box Sections

6.1 Box Sections

The trisection may also be used to create another class of configuration known as the 'box' or 'extended box' class [10]. The box section is similar to the cascade quartet section, that is, it has four resonator nodes arranged in a square; however the input to and output from the quartet are from opposite corners of the square. Fig. 13(a) shows the conventional quartet arrangement for a 4th degree filtering function with a single transmission zero and realized with a trisection. Fig. 13(b) shows the equivalent box section realizing the same transmission zero but without the need for the diagonal coupling. Application of the minimum path rule indicates that the box section can realize only a single TZ.

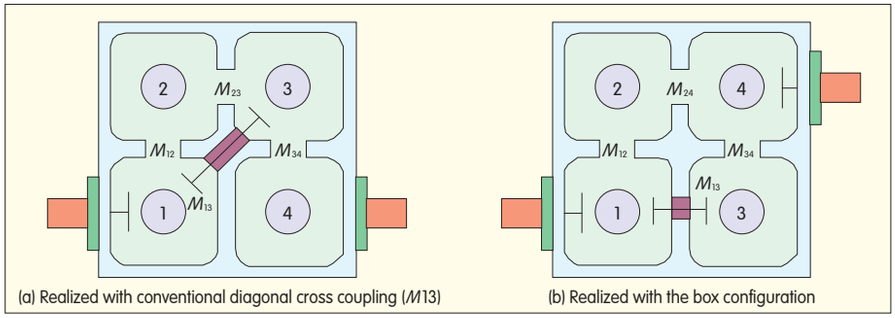
The box section is created by the application of a cross–pivot rotation to a trisection that has been synthesized within the overall coupling matrix for the filter. To transform the trisection into a basic box section, the rotation pivot is set to annihilate the second main–line coupling of the trisection in the coupling matrix. ie. pivot = [2,3] annihilating element M_{23} in the trisection 1–2–3 in the 4th degree example of Fig.13(a) and in its equivalent coupling and routing schematic in Fig.14(a). In the process of annihilating the main–line coupling M_{23}, the coupling M_{24} is created (Fig.14(b)), and then, by 'untwisting' the network, the box section is formed (Fig.14(c)).

In the resultant box section, one of the couplings is always negative, irrespective of the sign of the cross–coupling (M13) in the original trisection. Fig. 15(a) gives the coupling and routing diagram for a 10th degree example with two transmission zeros

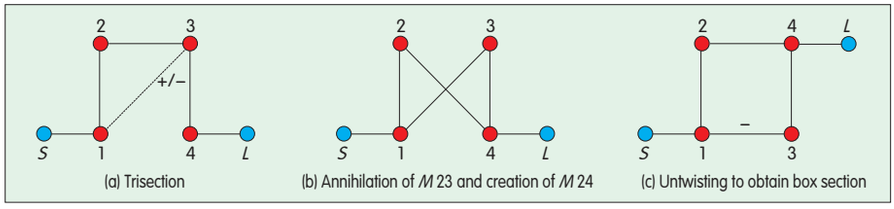


▲Figure 12. Transformation of 3 conjoined trisections to form a quintet section.

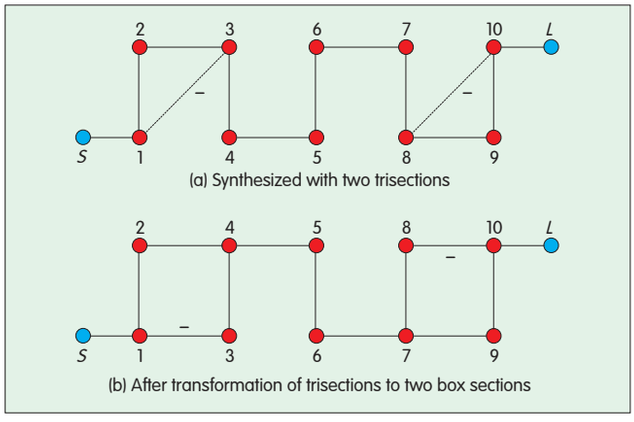
Richard J Cameron



▲Figure 13. 4–1 asymmetric filtering function.



▲Figure 14. 4–1 filter—formation of the box section.



◀Figure 15. 10–2 asymmetric filter–coupling and routing diagrams.

realized as trisections. Fig. 15(b) shows that each trisection has been transformed into a box section within the matrix by the application of two cross–pivot rotations at pivots [2], [3] and [8],[9]. Having no diagonal couplings, this form is suitable for realization in dual–mode technology.

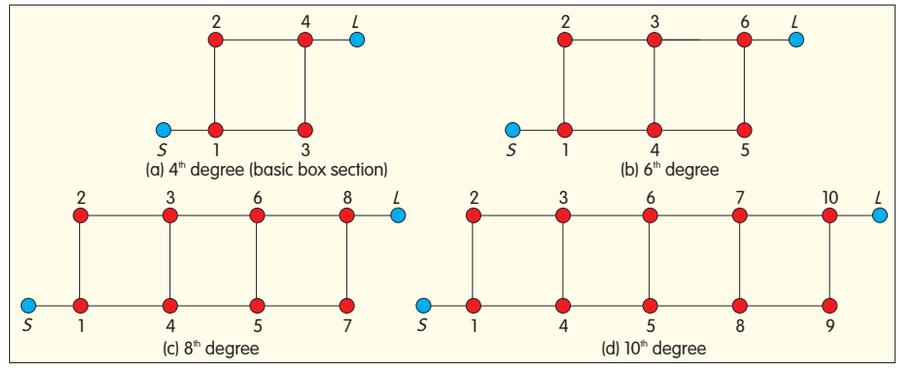
An interesting feature of the box section is that to create the complementary response (i.e. the transmission zero appears on the opposite side of the passband), it is only necessary to change the values of the self couplings to their conjugate values. In practice, this is a process of retuning the resonators of the RF device—no couplings need to be changed in value or sign. This means that the same physical structure can be used for the filters of, for example, a complementary diplexer.

6.2 Extended Box Sections

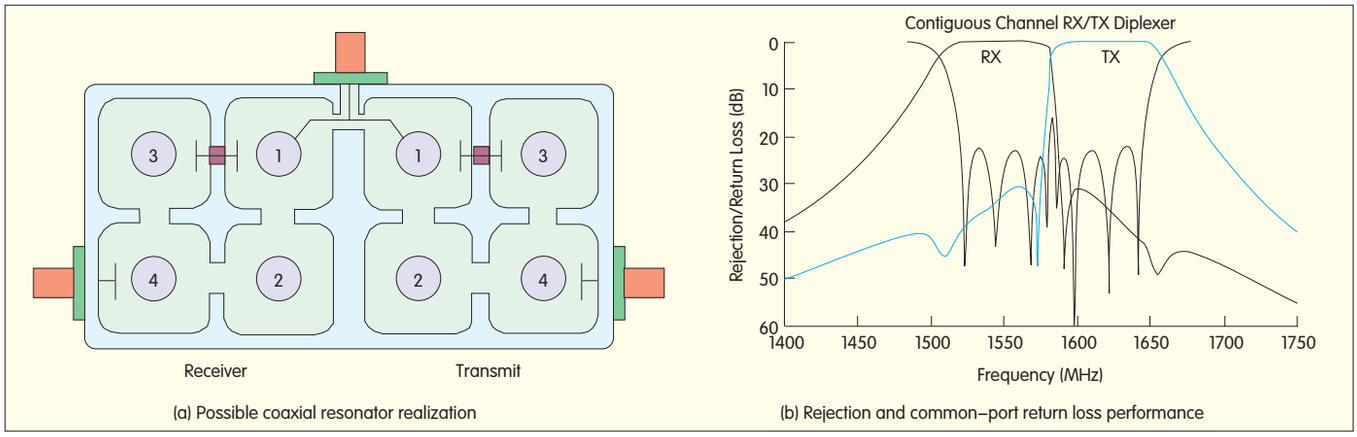
The basic box section may be extended to enable a greater number of transmission zeros to be realized, but retaining a convenient physical arrangement is shown in (Fig.16) [10]. Here, the basic 4th degree box section is shown and then the addition of pairs of resonators to form 6th, 8th and 10th degree networks. Application of the minimum path rule indicates that a maximum of 1, 2, 3, 4⋯ (N–2)/2 transmission zeros can be realized by the 4th, 6th, 8th , 10th,⋯Nth degree networks respectively. The resonators are arranged in two parallel rows with half the total number of resonators in each row. The input is at the corner of one end and output from the diagonally opposite corner at the other end. Even though asymmetric characteristics may be prescribed, there are no diagonal cross–couplings.

At present, a formal series of rotations to generate an extended box filter is not known. The form can, however, be derived with the software package Dedale–HF, which is accessible on the Internet [11].

Because of its simplicity, the box filter is useful for the design of transmit/receive diplexers, which are very often found in the base stations of cellular telephony systems. An example of a simple diplexer comprising two complementary–asymmetric 4th degree filters, each with one transmission zero producing a 30 dB rejection lobe over each other's usable bandwidth. Is shown in Fig. 17(a), and its performance is shown in Fig. 17(b). This diplexer was designed using software that optimizes the length and impedances of the common–port coupling wires as well as the first few elements of each filter nearest to the CP



▲Figure 16. Coupling and routing diagrams for extended box section networks.



(a) Possible coaxial resonator realization

(b) Rejection and common–port return loss performance

▲Figure 17. Transmit/receive diplexer for base station applications

(coupling values, resonator tuning frequencies). In practice, much greater Tx–Rx isolation is usually required, and higher degree filters with more transmission zeros have to be used.

7 Conclusions

In this article some of the more recent developments in the art of filter synthesis have been outlined. These have been based on the coupling matrix representation of the filter network's inter–resonator coupling arrangements because of the amenity of the coupling matrix to mathematical manipulation, and the one–to–one correspondence of the elements of the coupling matrix to the real filter parameters.

The methods described in this article probably do not cover all those available today for filter network synthesis. Some configurations cannot be achieved by a sequence of analytically–calculated rotations, and optimization methods working on the coupling matrix elements have to be employed [12]–[13]. Some advanced developments are ongoing into the synthesis of 'lossy' filters [14], which are used to compensate for a low resonator Q and give very linear in–band performance but at the expense of high–ish insertion loss (not a real problem in low–power circuits). Also, some work is also ongoing into the synthesis of coupling matrices for wideband devices where the coupling elements have a frequency dependency [15]. Some novel synthesis techniques have recently become available for the design of circuits incorporating the non–resonant node (NRN) element, which are useful in high–power applications and for making the design of dielectric and planar circuits easier [16].

References

[1] S. Darlington, "Synthesis of reactance 4–poles which produce insertion loss characteristics," *J. Math. Phys.*, vol. 18, pp. 257–353, 1939.
[2] M. E. van Valkenburg, *Network Analysis*. Englewood Cliffs, N.J.: Prentice–Hall, 1955.
[3] G. Matthaei, L. Young, and E. M. T. Jones, *Microwave Filters, Impedance Matching Networks and Coupling Structures*. Norwood, MA: Artech House, 1980.
[4] J. D. Rhodes, "The generalized direct–coupled cavity linear phase filter," *IEEE Trans. Microw. Theory Tech.*, vol. 18, no. 6, pp. 308–313, June 1971.
[5] A. E. Atia and A. E. Williams, "New types of bandpass filters for satellite transponders," *COMSAT Technical Review*, vol. 1, no.1, pp. 21–43, 1971.
[6] R. J. Cameron, "General coupling matrix synthesis methods for Chebyshev filtering functions," *IEEE Trans. Microw. Theory Tech.*, vol. 47, no. 4, pp. 433–442, Apr. 1999.
[7] R. J. Cameron, "Advanced coupling matrix synthesis techniques for microwave filters," *IEEE Trans. Microw. Theory Tech.*, vol. 51, no. 1, pp. 1–10, Jan. 2003.
[8] H. C. Bell, "Canonical asymmetric coupled–resonator filters," *IEEE Trans. Microw. Theory Tech.*, vol. 30, no. 9, pp. 1335–1340, Sept. 1982.
[9] S. Tamiazzo and G. Macchiarella, "An analytical technique for the synthesis of cascaded N–tuplets cross–coupled resonators microwave filters using matrix rotations," *IEEE Trans. Microw. Theory Tech.*, vol. 53, no. 5, pp. 1693–1698, May 2005.
[10] R. J. Cameron, A. R. Harish, and C. J. Radcliffe, "Synthesis of advanced microwave filters without diagonal cross–couplings," *IEEE Trans. Microw. Theory Tech.*, vol. 50, no. 12, pp. 2862–2872, Dec. 2002.
[11] Dedale–HF page. [Online]. Available: http://www.sop–inria.fr/apics/Dedale
[12] S. Amari, "Synthesis of cross–coupled resonator filters using an analytical gradient–based optimization technique," *IEEE Trans. Microw. Theory Tech.*, vol. 48, no. 9, pp. 1559–1564, Sept 2000.
[13] W. A. Atia, K. A. Zaki, and A. E. Atia, "Synthesis of general topology multiple–coupled resonator filters by optimization," in *IEEE MTT–S Int. Microw. Symp.* vol. 2, Baltimore, MD, 1998, pp 821–824.
[14] V. Miraftab and M. Yu, "Advanced coupling matrix and admittance function synthesis techniques for dissipative microwave filters," *IEEE Trans. Microw. Theory Tech.*, vol. 57, no. 10, pp. 2429–2438, Oct. 2009.
[15] J. Rhodes and I. C. Hunter, "Synthesis of reflection–mode prototype networks with dissipative circuit elements," *IEEE Proc. Microw., Antennas, Propag.*, vol. 144, no. 6, pp. 437–442, Dec. 1997.
[16] S. Amari, F. Seyfert, and M. Bekheit, "Theory of coupled resonator microwave bandpass filters of arbitrary bandwidth," *IEEE Trans. Microw. Theory Tech.*, vol. 58, no. 8, pp. 2188–2203, Aug. 2010.
[17] S. Amari and U. Rosenberg, "New building blocks for modular deisgn of elliptic and self–equalized filters," *IEEE Trans. Microw. Theory and Techniques.*, vol. 52, no. 2, pp. 721–736, Feb. 2004.

Biographies

Richard J. Cameron CEng. FIET, FIEEE received the B.Sc. degree in telecommunications and electronic engineering from Loughborough University, U.K., in 1969. Later that year, he joined Marconi Space and Defence Systems Company in the U.K., where his activities included small earth–station design, telecommunication satellite system analysis, and computer–aided RF circuit and component design. In 1975, he joined the European Space Agency's technical establishment (ESTEC, The Netherlands), where he was involved in the research and development of advanced microwave active and passive components and circuits, with applications in telecommunications, scientific and earth observation spacecraft. Since joining Com Dev Ltd. in 1984, he has been involved in the software and methods for the design of high–performance components and sub–systems for both space and terrestrial application.

Richard Cameron is now retired, but still performs duties as a Visiting Professor at Leeds University (UK), and is active in writing technical articles, presenting lectures, examining PhD candidates and providing consultancy services on an ad hoc basis.

Privacy–Preserving Protocol for Data Stored in the Cloud

Abstract: Data storage is an important application of cloud computing. With a cloud computing platform, the burden of local data storage can be reduced. However, services and applications in a cloud may come from different providers, and creating an efficient protocol to protect privacy is critical. We propose a verification protocol for cloud database entries that protects against untrusted service providers. Based on identity–based encryption (IBE) for cloud storage, this protocol guards against breaches of privacy in cloud storage. It prevents service providers from easily constructing cloud storage and forging the signature of data owners by secret sharing. Simulation results confirm the availability and efficiency of the proposed protocol.

Keywords: privacy; cloud storage; IBE; secret sharing

Hongyi Su
Geng Yang
Dawei Li
(College of Computer Science, Nanjing University of Posts and Telecommunication)

Cloud computing is a new service platform. When data is stored in clouds, the owner no longer has physical possession of the data's storage. But untrusted service providers have independent administrator authority over the data, and this creates potential security threats. Data integrity is also a security challenge in cloud computing [1]. Protecting the privacy of data owners has become a new challenge in cloud computing [2].

In this paper, we propose a privacy–preserving protocol with elliptic curve cryptography and secret sharing. In a given scenario, service providers comprise coordinate servers in which sensitive data (entries of cloud databases) are re–computed with k servers (k is a threshold value). Users also verify the shares from k servers on the condition that they (the users) only rely on the trust of data owners' signatures.

The work is supported by the National Natural Science Foundation of China under Grant No. 60873231, the National Basic Research Program of China ("973" Program) under Grant No. 2011CB302903, the High Education Natural Science Foundation of Jiangsu Province under Grant No. 08KJB520006 and Funds of Key Lab of Fujian Province University Network Security and Cryptology under Grant No. 09A010, and Innovation Project for postgraduate cultivation of Jiangsu Province, China under Grant No. CX10B_195Z.

1 Related Work

Identity–Based Encryption (IBE) is a key technique for the proposed protocol. A novel IBE on a bilinear map was proposed in [3]. However, it failed in a distributed environment and did not prevent collusion between k dishonest service providers. Joonsang Baek [4] constructed the first identity–based threshold decryption, which is secure against chosen–ciphertext attack.

Baek's scheme is based on the concept of "federal identity," introduced by Liang Yan [5], and is a new solution to strengthening cloud security.

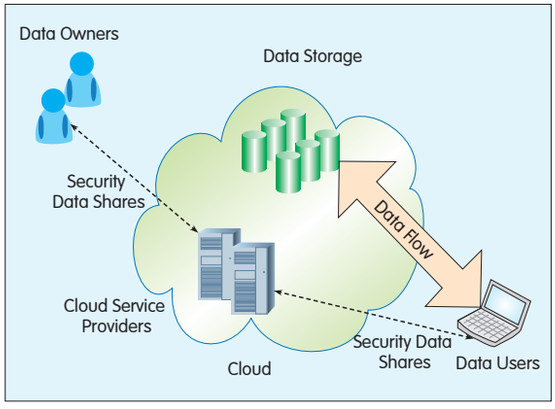
Brian Thompson [6] leveraged database entries of data owners as shared secrets. Owner identities can be replaced by database entries and used as owners' signatures. This can prevent disclosure of private information from the identity. Cong Wang [2] introduced a third party auditor (TPA) to audit cloud service providers before a user accesses cloud storage services. TPA guarantees that data integrity will be maintained and service providers will behave ethically.

We use database entries as federal identities to verify service providers' data. We also use threshold encryption on a bilinear map to adapt to distributed cloud storage.

2 The Proposed Model and Protocol

2.1 The Proposed Model

The model we propose aims to protect cloud data against untrusted service providers. This model involves data owners, cloud service providers, and data users, as shown in Fig. 1. Data owners store data in the cloud and send every share of data entries to k service providers. Cloud service providers collaboratively store data, especially the shares of cloud databases entries. They have the signature of database entries (shadow secret) from data owners. Data users access data from the service providers and have access to the public information of data owners in order to



▲Figure 1. The proposed cloud storage model.

verify the shares received from service providers.

2.2 Process of The Proposed Protocol

The proposed protocol is based on IBE [3] and secret sharing [7]. The identity in this protocol is replaced by the entry of cloud databases as a secret to be shared. Strings or multimedia data can be converted by hash functions. A data owner outsources the database (such as m shares of the entry for the database) to m service providers and requires that at least any k servers can reconstruct the entry of the database collaboratively. However, $k-1$ servers cannot. Shadow secrets are also proposed to verify each share of the entry. Furthermore, the data owner and data user can delegate one of m providers as the primary provider SP_u in order to reduce the cost of communication. SP_u gathers other $k-1$ shares and shadow secrets and sends k shares to users. This solution requires that should be absolutely trustworthy. Finally, the primary provider (delegated by data users) reconstructs the entry of the database and verifies the shares of other $k-1$ providers with shadow secrets.

The proposed protocol is implemented in four phases: initialization, distribution, verification, and reconstruction.

(1) Initialization

The proposed model takes a security parameter h and chooses a big primer p (h bit) in order to find a hypersingular elliptic curve E/GF(p). The following notations are used:

- l is the length of plaintext, and we use hash function to convert entry x into plaintext
- P is the generator of subgroup $(G,+)$ in E/GF(p)
- q ($q>2^h$) is the order of (G, +)
- G and G* are two (multiplicative) cyclic groups of prime order p
- Z_q^* denotes the group $\{0; \ldots; q-1\}$ under addition modulo q
- $\hat{e}$: is a bilinear map $\hat{e}$: $G\times G \rightarrow GF(P^2)^*$
- H_1, H_2 are one–way hash functions H_1: $\{0,1\}^* \rightarrow G^*$; H_2: $GF(P^2)^* \rightarrow \{0,1\}^{\prime}$.

First, a master key s is chosen to compute system public key $P_{pub}=sP$. Broadcast system parameters $\{G, q, P, \hat{e}, H_1, H_2, P_{pub}\}$ are also set for data owners, data users, and service providers.

(2) Distribution

The data owner distributes m shares of the entry E of the database to m service providers along with the verification key. Let E be the entry of database, and use IBE to produce plaintext(c,u), where

$$C=E\oplus H_2(\hat{e}(D_e, U)); U=rP, r\in Z_q^* \quad (1)$$

The process of distributing shares has the following six steps :

Step1: Execute the function Extract from IBE [3] to compute a key pair $<Q_E, D_E>$, $Q_E=H_1(E)$; $D_E=sQ_E$

Step2: Randomly choose k coefficients from $G^*\{a_j | a_j \in G^*; j=1,2,\cdots, k-1; a_{k-1}\neq 0\}$ and compue m shares of De with a polynomial $F(x)$ of order $k-1$,

$$F(x)=D_E+\sum_{j=1}^{k-1} a_j x^j \quad x\in\{0\}\cup \mathrm{IN}$$

Step3: Compute service provider SP_i's shadow secret $S_i=F(i)$, then compute $y_i=e(S_i, P)$, $1\leqslant i\leqslant m$

Step4: Compute verification key $U_j=\hat{e}(a_j, P_{pub})$, $1\leqslant j\leqslant t-1$, $U_0=\hat{e}(D_{ID}, P_{pub})$

Step5: Send S_i to SP_i with a private channel, and broadcast y_i, U_j; $1\leqslant i\leqslant m$, $0\leqslant j\leqslant k-1$

Step6: Compute Proof for verification key.

After choosing $w_i\in(G, +)$, data owners compute $E_{1i}=\hat{e}(w_i, P_{pub})$, $E_{2i}=\hat{e}(w_i, P)$, $c_i=H_2(E_{1i}+E_{2i})$, $r_i=w_i-S_i c_i \bmod q$. Owners broadcast $PROOF_i=(r_i, c_i)$, $1\leqslant i\leqslant m$.

(3) Verification

SP_μ applies for other $k-1$ providers' shadow secrets S_i and uses the public information to verify the validity of S_i in the following steps:

(a) SP_μ collects $U_i (0\leqslant i\leqslant k-1)$, $y_i (1\leqslant i\leqslant m)$ to compute $Y_j=\prod_{l=0}^{t-1} U_l^{j^l}$

(b) It verifies $E_{1i}=\hat{e}(r_i, P_{pub})Y_i^{c_i}$, $E_{2i}=\hat{e}(r_i, P)Y_i^{c_i}$ and then $H_2(E_{1i}+E_{2i})$ to compare with c_i from $PROOF_i=(r_i, c_i)$. If the result is equal, S_i is proven valid.

(4) Reconstruction

After verification, SP_μ collects S_i from other $k-1$ service providers. Given the ciphertext $<C, U>$, SP_μ uses the following expression to revert to the entry E:

$$sum=\prod_{j\in\Phi}\hat{e}(s_i, U)^{C_{0j}^{\Phi}} \quad (2)$$

$$C_{x,j}^{\Phi}=\prod_{t\in\Phi, i\neq j}\frac{x-l}{j-l}\in Z_q, \quad \Phi\subset\{1,2,\cdots,n\}, |\Phi|\geqslant t \quad (3)$$

Then following IBE, SP_μ reduces the entry E:

$$E=C\oplus H_2(sum) \quad (4)$$

2.3 Verification of the Proposed Protocol

The proposed protocol is proven correct by calculating Y_i, given U_i $(0\leqslant i\leqslant k-1)$.

(1) Verification of the validity of shadow secret S_i

The primary provider SP_m collects information set $(P_{pub}, U_i; P, y_i; S_i)$ for the verification.

The correctness of $Y_i=\prod_{j=0}^{t-1} U_j^{i^j}$ is given below:

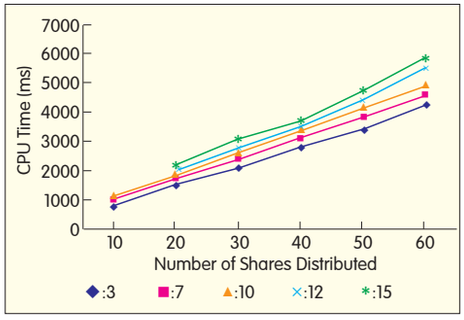
$$Y_i=\prod_{j=0}^{t-1} U_j^{i^j}=\hat{e}(D_{ID}+\sum a_j i^j, P_{pub}) =\hat{e}(S_i, P_{pub}) \quad (5)$$

According to Y_i, y_i, it is easy to prove the correctness of $PROOF_i=(r_i, c_i)$ as follows:

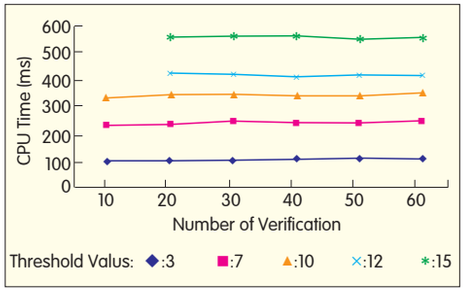
$$E_{1i}=\hat{e}(r_i, P_{pub})Y_i^{c_i} =\hat{e}(w_i-S_i c_i, P_{pub})\hat{e}(S_i, P_{pub})^{c_i} =\hat{e}(w_i, P_{pub}) \quad (6)$$

$$E_{2i}=\hat{e}(r_i, P_{pub})Y_i^{c_i} =\hat{e}(w_i-S_i c_i, P)\hat{e}(S_i, P)^{c_i} =\hat{e}(w_i, P) \quad (7)$$

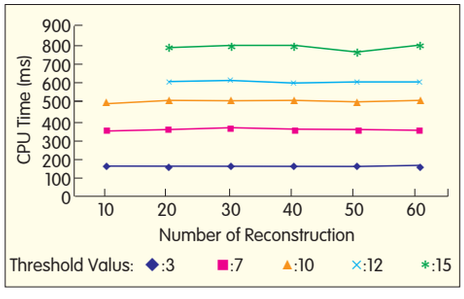
With the computation of c_i in $PROOF_i=(r_i, c_i)$, $H_2(E_{1i}+E_{2i})$ is equal to c_i. $PROOF_i=$



▲Figure 2. Comparison on time cost of distributing shares.



▲Figure 3. Time cost of verification.



▲Figure 4. Time cost of reconstruction.

(r_i, c_i) can prove the validity of S_i.

(2) Reconstruction of the entry (E) of the database:

After verifying k (threshold value) shadow secrets S_i, the primary provider computes a bilinear map with U in the ciphertext and S_i. The correctness is proven as follows:

$$sum=\prod_{j\in\Phi}\hat{e}(S_j,U)^{C^{\Phi}_{0,j}}$$
$$=\hat{e}(\sum_{j\in\Phi}C^{\Phi}_{0,j}S_j,U)=\hat{e}(D_e,U) \quad (8)$$

According to IBE, SP_μ reverts the entry $E=C\oplus H_2(\text{sum})$.

3 Simulation

The simulation was conducted using C language on a Windows XP system running on two 2.0 GHz Intel Core processors each with 2 GB of RAM and a Western Digital 320 GB Serial ATA driver. Multiprecision Integer and Rational Arithmetic C/C++ Library (MIRACL) version 5.3 was used. The simulation result is the mean of ten trials.

(1) Cost of Distributing Shares of Secret and Signature by Data Owners

We assess the performance of distributing shares according to two criteria: CPU time and the number of shares distributed. As shown in Fig.2, when the total number of service providers increases (based on different threshold values), the time cost increases. When the threshold value increases, the time cost also increases. So when a data owner distributes shares of database entry and signature, the total number of service providers, as well as the threshold value, needs to be considered. Moreover, in order to guarantee protection against collusion from service providers, the range of the threshold should be considered. A threshold value of half the total number of service providers is a good choice for balancing security and efficiency.

(2) Comparison of the Time Costs of Distribution, Verification, and Reconstruction by Data Users

The time cost of reconstruction and of verification are similar at the same threshold level regardless of the total number of service providers, as shown in Figs. 3 and 4. In the proposed protocol, verification and reconstruction are run by the primary service providers delegated by data users. So the primary service provider can save communication cost by gathering k–1 shares. The figures show that efficiency in verification and reconstruction is greater than that in distribution.

4 Conclusion

In this paper, we propose a privacy–preserving protocol for data storage in the cloud. We focus on stopping data being disclosed by untrusted service providers when data owners distribute their database entries. By using IBE algorithm and a secret–sharing scheme, the protocol protects against collusion by multiple service providers. The simulations show that regardless of network environment, the proposed protocol is correct and efficient.

References

[1] W. Zeng, Y. Zhao, O. Kairi, and W. Song, "Research on cloud storage architecture and key technologies, " in *Proc. 2nd Int. Conf. Interaction Sciences*, Seoul, 2009, pp.1044–1048.
[2] Q. Wang, C. Wang, K. Ren, and W. Lou, "Privacy–preserving public auditing for data storage security in cloud computing," in *Proc. IEEE INFOCOM*, San Diego, CA, 2010, pp.1–9.
[3] D. Boneh, B. Lynn, and H. Shacham, "Short signatures from the weil pairing," *J. Cryptology*, vol.17, no.4, pp. 297–319, 2004.
[4] J. Baek, and Y. Zheng, "Identity–based threshold decryption," in *Proc. 7th Int. Workshop on Public Key Cryptography (PCK 2004)*, Singapore, pp.262–276.
[5] L. Yan, C. Rong, and G. Zhao, "Strengthen cloud computing security with federal identity management using hierarchical identity–based cryptography," in *Proc. 1st Int. Conf. Cloud Comput. (CloudCom'09)*, Beijing, pp. 167–177.
[6] B. Thompson, S. Haber, W.G. Horne, and T. Sander, D. Yao, "Privacy–preserving computation and verification of aggregate queries on outsourced databases," in *Proc. 9th Int. Symp. on Privacy Enhancing Tech.(PETS'09)*, Seattle, WA, pp.185–201.
[7] A. Shamir, "How to share a secret," in *Commun. ACM*, vol.22, no.11, pp.612–13, 1979.
[8] A. Shraer, C. Cachin, A. Cidon, I. Keidar, Y. Michalevsky, and D. Shaket, "Venus: Verification for untrusted cloud storage," in *Proc. ACM Workshop on Cloud Comput. Security (CCSW'10)*, Chicago, Ill., pp.19–30.
[9] E. Bertino, F. Paci, and R. Ferrini, "Privacy–preserving digital identity management for cloud computing," *IEEE Comput. Soc. Data Engineering Bulletin*, pp. 1–4, 2009.
[10] W. Itani, A. Kayssi, and A. Chehab, "Privacy as a Service: Privacy–Aware Data Storage and Processing in Cloud Computing Architectures," in *Proc. 8th IEEE Int. Conf. on Dependable, Autonomic and Secure Comput.*, Chengdu, China, 2009, pp.711–716.

Biographies

Hongyi Su (noman_michael@163.com) is a master's candidate at the College of Computer Science, Nanjing University of Posts and Telecommunications. His research interests include information security and computer science.

Geng Yang (yangg@njupt.edu.cn) is a professor and doctoral advisor at the College of Computer Science, Nanjing University of Posts and Telecommunications. He received his Ph.D. degree in computer science from Laval University, Canada. Professor Yang is a member of the IEEE Computer Society and a Standing Member of Chinese Computer Education Society. His research interests include network security, parallel and distributed computing, and mobile computing.

Dawei Li (lidw1981@163.com) is a Ph.D. candidate at the College of Computer Science, Nanjing University of Posts and Telecommunications. His research interests include computer networks and information security.

A Mobility Management Solution Based on ID/Locator Separation

Abstract: Current mobility management solutions based on ID/Locator separation are not easily deployed and cannot solve routing scalability and mobility problems. This paper proposes a novel network architecture based on ID/Locator separation and suggests a new mobility management solution. This solution solves the problem of scalability in the network and also provides better support for mobility. It can be easily deployed because no modification of the mobile host's protocol stack is required. The identifier contains some routing information; so the solution provides intrinsic interworking with traditional mobile hosts. Because the mapping systems are distributed to the edge networks, robustness of the whole system is enhanced and handover delay is decreased.

Keywords: identifier and locator separation; mobility management; location management; handoff control

Yuhong Li[1]
Yunjing Hou[2]
Shiduan Cheng[1]
(1.State Key Laboratory of Networking and Switching Technology, Beijing University of Posts & Telecommunications;
2. Datang Wireless Mobile Innovation Center, China Academy of Telecommunication Technology)

The basic TCP/IP protocols of the Internet are designed for static hosts, and an IP address bears the semantics of locator and identifier. So the Internet does not, by nature, support mobility. To address the mobility issue, the IETF has proposed several mobility management protocols based on IP. These include mobile IP (MIP) [1],[2] and proxy MIP version 6 (PMIPv6) [3]. These protocols enable the Internet to support moving hosts. But they cannot solve the problem of routing scalability caused by the double semantics of IP addresses [4]–[6].

To solve these problems, the basic approach is to decouple the dual semantics of IP addresses. Some renowned institutes have conducted research on this issue and put forward solutions. These solutions include Host Identity Protocol (HIP) [7], by R. Moskowitz et al.; SHIM6 [8] by Nordmark et al.; Locator/ID Separation Protocol (LISP) [9] by Cisco; Six/One [10] by Christian Vogt; and global locator, local locator, and identifier split (GLI–split) [11] by Michael Menth et al. On top of the network architecture based on ID/locator separation, traditional mobility management protocols have poor scalability. The centralized deployment of mobility management entities makes these entities prone to single point failure. Moreover, they are likely to become bottlenecks for the data communications of mobile hosts. Therefore, traditional mobility management solutions are no longer suitable, and a new solution is needed for network architecture based on ID/locator separation.

This paper analyzes the disadvantages of existing mobility management solutions based on ID/locator separation. It proposes a novel network architecture and a new mobility management solution. This new solution is designed to address the mobility problem by overwriting the destination address.

This work is funded by the European Commission funded ICT–FP7 IP Project EFIPSANS under Grant No. INFSO–ICT–215549, and the National Basic Research Program of China ("973" Program) under Grant No. 2009CB320504.

1 Existing Mobility Management Solutions Based on ID/Locator Separation

Existing mobility management solutions based on ID/locator separation fall into three categories: host–based, network–based, and host and network–based.

HIP is a typical host–based solution. In HIP, a host identity layer is added between the transport layer and the IP layer to decouple the dual semantics of an IP address. The transport layer and its upper layers are shielded from any change of the host IP address. But deploying HIP is difficult because it requires changing the host and deploying a large number of RendezVous servers (RVSs) [12] in the network. Moreover, HIP does not support broadcast services. When the two communicating parties move at the same time, a long handover delay is incurred. The SHIM6 protocol is another host–based solution. It divides the IP

layer into three sublayers. Among them, the SHIM sublayer maintains the association between the identifiers and the locators of the two hosts in each session. These sublayers also shield the upper layers from any change of the hosts' locators. However, detecting available address pairs may introduce a long delay; and thus, it provides poor mobility support.

LISP is a network–based solution that only requires enhancement of the functions of edge routers in edge networks. No modification of the host is required. However, LISP is a protocol based on encapsulation. The data packets from a host must be encapsulated by the ingress tunnel router (ITR) before they are sent to the core network, which increases the bandwidth consumption. In LISP, each host is assigned a unique endpoint identifier (EID) in each edge network. When a host moves from one edge network to another during communication, the change of EID interrupts the TCP connection. To address this problem, LISP mobile node (LISP MN) architecture [13] is proposed. However, this architecture also has also some disadvantages. First, before deploying LISP MNs, functions of the hosts have to be enhanced. Second, an LISP MN can directly access the mapping system, which may cause some security problems for the entire system. Third, to enable interconnection between LISP MNs and traditional hosts, a proxy egress tunnel router (PETR) is needed in the network. Fourth, a mobile host does not have a mobile anchor point in the network. So data packets sent from the correspondent host (CH) to the mobile host in the period between when the mobile host starts to move and when the CH obtains the new IP address of the mobile host will be lost. A Six/One router is a network–based solution that uses an address overwriting method. The edge network is connected to the core network via Six/One routers. A host is assigned two addresses: a unique endpoint address in the edge network and a transmission address for global routing. The two addresses are one–to–one mapped, but the transmission overhead is quite large. Communication between an enhanced host and a traditional host is likely to be interrupted when the enhanced host moves.

GLI–split architecture is a solution based on both host and network. It divides an IP address into three address spaces: global address, local address, and identifier address. On the host side, a vertical address translation function is introduced. It translates the identifier address used in the transport layer into a local or global address. The host uses the identifier address to set up a communication association. However, the two–level mapping system in GLI–split may cause a large handover delay. In addition, the host is allowed to access the mapping system, which may cause some security problems.

In brief, existing mobility management solutions based on ID/locator have potential security problems, are difficult to deploy, and may have long handover delay. Therefore, it is necessary to suggest a new network architecture and develop a mobility management mechanism suitable to the architecture.

2 A New Mobility Management Solution Using Destination Address Overwriting

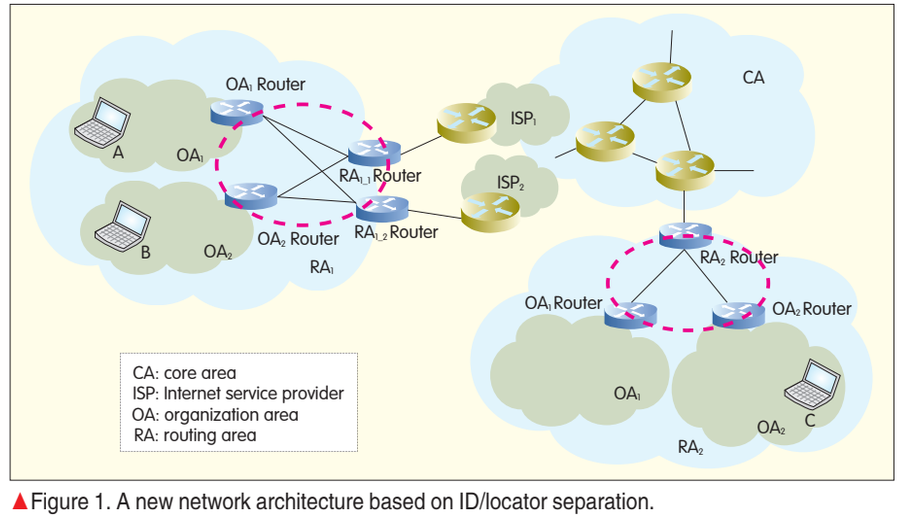
An ideal mobility management solution based on ID/locator separation should:

- be compatible with traditional Internet so that deployment is less difficult and modifications to the host and Internet are kept to a minimum
- allow only network entities to access the mapping system to ensure system security
- have the mapping systems deployed close to or within an edge network to shorten handover delay
- have the mapping systems that store mapping between the ID and the locator of a host deployed in a distributed way in order to provide robustness
- have the address space used by the edge network separate from that used by the core network to make the network more scalable
- support various types of applications, including unicast, broadcast, and multicast applications
- support hosts with multiple interfaces so that multihoming services can be delivered.

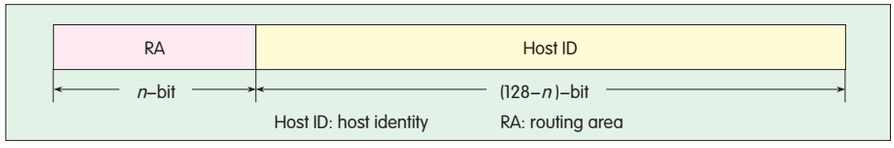
Taking into account these characteristics, this paper proposes a new network architecture based on ID/locator separation with a mobility management solution.

2.1 New Network Architecture Based On ID/Locator Separation

Fig. 1 illustrates the new network architecture based on ID/locator separation. In this architecture, the network is divided into core area (CA) and routing area (RA). The CA has the



▲Figure 1. A new network architecture based on ID/locator separation.



▲Figure 2. Identifier format.

same functions as those of existing Internet backbone and consists of high–speed routers. One RA is made up of several organization areas (OAs).

RA is edge network, and its range depends on the specific deployment. For example, an RA may be constructed according to geographical location. An OA is often related to an organization—one company can be an OA. Each RA is connected to the CA by one or several RA routers.

The source and destination IP addresses of a packet are the IDs of the source and destination host respectively. The default gateway of the source host is configured as the address of the OA router the host is currently connected to. When an OA router receives a packet, it looks for the mapping between the ID and the locator in the local buffer based on the ID of the destination host. If a mapping record is found, the OA router forwards the packet within the OA. Otherwise, it uses the global routable IP address to overwrite the destination IP address field of the packet. First, it looks in the mapping system for the global routable address of the destination host according to the ID of the destination host and caches the result in the local buffer. Then, it uses this global routable address to overwrite the destination IP address field of the packet. After overwriting, the OA router sends the packet out. When the OA router of the destination host (denoted as OA–router–D) receives the packet, it looks up the ID of the destination host in the <ID, global locator> mapping stored locally according to the destination IP address. It then uses the found ID to overwrite the destination address field and forwards the packet to the destination host.

2.2 ID and Locator

The ID is the unique information used to identify a host. So that the protocol stack and applications of the host are not modified, the IDs in the new network architecture are designed to have the same form as an IP address. The length of the ID is the same as an IPv6 address, that is, 128 bits. The ID is made up of two parts: RA information and host ID. The RA information is the prefix of the router that stores the host's mapping information (denoted as Router–M). Its length is n bits. The host ID is the globally unique identity of the host. This ID can be generated, for example, in the same way as a host identifier tag (HIT) in HIP. However, only 128–n bits of the Hash value are truncated. The format of the ID is shown in Fig. 2.

The solution proposed here is based on IPv6, so we assume that all traditional hosts support IPv6 protocol stack, and we call the host registered in the ID/locator separation–based system the extended host.

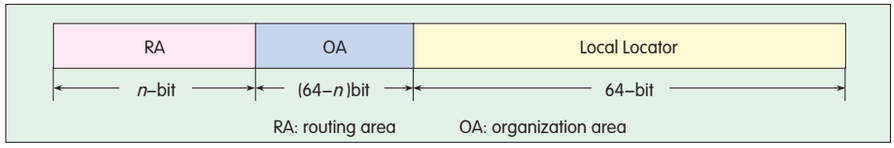
Current solutions that use source and destination address overwriting have two problems in terms of interconnection between the extended host and the traditional host. One is that the global routable addresses of the extended hosts are stored in the domain name server (DNS). Because the DNS is updated at relatively long intervals, a traditional host may obtain the old IP address of an extended host. Consequently, the traditional host cannot initialize communication with the extended host. The other problem is that the session between an extended host and traditional host may be interrupted when the extended host moves and the traditional host fails to obtain the new global routable address of the extended host.

Our solution to the first problem is to record the ID of the extended host in the DNS. Because the ID of the extended host is static, the connection establishment problem caused by slow DNS update can be avoided. Our solution to the second problem is to include some routing information in the ID and to adopt a communication mode with optimized routing based on proxy. Regardless of whether a session is initialized by an extended host or a traditional host, the source address of the first packet sent from the traditional host to the extended host is the global routable address of the traditional host; the destination address is the ID of the extended host. Because some routing information is included in the ID, the first packet will finally be routed to RA–Router–M. When the extended host moves from OA router 1 to OA router 2 during communication, OA router 1 is responsible for updating the location information of the extended host recorded at the correspondent host. If the correspondent host is an extended one, OA router 1 sends a Transfer message to the OA router of the correspondent host E (more details can be found in section 2.4). Otherwise, OA router 1 sends a normal IPv6 binding update message to the traditional host, which, in turn, processes the message according to MIPv6. Therefore, by including routing information in the ID and optimized proxy routing, our solution provides better backward compatibility.

The locator is the global routing identity of a host, and it changes with the location of the host. It is a 128–bit IPv6 address consisting of RA, OA, and local locator. RA + OA is the prefix of the OA router to which the host connects. The local locator is a local address assigned by the OA router and is only valid within the coverage of the OA router. A host can only "see" its local locator. The format of the locator is shown in Fig. 3.

2.3 Mapping System

The mapping system stores the relationship between the locator and the ID. It is comprised of local mapping systems within different RAs. To avoid introducing new entities to the network, the mapping system can be deployed on routers. To guarantee the robustness of the mapping system, the distributed hash table (DHT) is used to construct an overlay using routers



▲Figure 3. Locator format.

within RAs.

When a host is registered in the system for the first time, the local OA router stores the mapping between its ID and locator in the mapping system of the local RA. The router in which the mapping information is stored is referred to as Router–M. The selection of Router–M depends on the DHT protocol being used.

When a host moves to another OA, the new OA router is responsible for updating the ID/locater mapping of the host in the mapping system. Because the RA field in the ID of the host contains the prefix of the RA where the Router–M is located, the new OA router can find the corresponding Router–M and update the ID/locater mapping of the host.

2.4 Mobility Management Solution

The mobility management solution here involves location management and handover control. Three signaling processes are included: registration, location update, and handover control.

2.4.1 Registration Process

Registration occurs when a host joins the system for the first time. The local OA router registers the ID/locator mapping information of the host into the mapping system.

The OA router periodically broadcasts a router advertisement message. When a host enters the coverage of the OA router, it generates its own ID based on the router advertisement message and sends a Register message to the OA router. The parameter of the Register message is the ID of the host, ID–H. After receiving the Register message, the OA router assigns a local locator to the host and saves the mapping <ID–H, local locator> in local buffer. Then it performs a DHT operation and stores the mapping information of the host in the local mapping system of the local RA. Consequently, the host–related information, including the locators of both the host (global–locator–H) and the OA router (global–locator–OA–Router), is stored in Router–M. When the OA router receives the Put ACK message, it returns the Register ACK message to the host, indicating that registration is successful.

2.4.2 Location Update Process and Handover Control Process

Location update means the update of the mapping system on account of a change of the locator. This change is caused by movement of the host during communication with a CH. For optimized routing, the location update process should update information of the host in the mapping system as well as the host's locator cached in the OA router of the CH.

A host may move in different scenarios. Here, we only take for example the most complicated case to describe the location update and handover control process. The host moves from the coverage of OA router 1 to the coverage of OA router 2 when it communicates with one CH. OA router 1 and OA router 2 are in different RAs, denoted as old RA and new RA respectively. Router–M of the host is in neither the old RA nor the new RA.

The location update and handover control process is as follows:

(1) The host moves into the coverage of OA router 2 and receives the advertisements broadcast by OA router 2.

(2) After receiving the router advertisement message, the host knows that it has entered the coverage of a new OA router and sends a Register message containing its ID to OA router 2.

(3) After receiving the Register message from the host, OA router 2 assigns a local locator to the host and generates a global routable address (global–locator–H–new) for the host. Then it sends an Update message to RA router 2 requesting to update the location information of the host in the mapping system. In the update message, the host's ID and new locator are included.

(4) When receiving the Update message, RA router 2 determines which RA the host's Router–M is located in. RA router 2 determines this location based on the RA information in the host's ID and then sends an Update message to the RA router.

(5) Once the RA router (of the RA where the host's Router–M is located) receives the Update message, it triggers the DHT update process within the local RA.

(6) Router–M then receives the Update message. It looks up the record of the host in the local buffer and sends a Forward message to OA router 1—where the information of the host has been updated with the new global routable address of the host and global routable address of OA router 2. The parameters of the Forward message include the host's ID–H and new locator.

(7) After receiving the Forward message, OA router 1 stores the parameters of the message in local buffer. When it receives a data packet destined for the ID of the host, it uses the new global routable address of the host to overwrite the destination IP address field. Because OA router 1 has the mapping relation between the ID and the locator of the CH, it sends a Transfer message to the OA router that the CH connects to and requests to forward the data packet destined for the host to the new global routable address of the host.

2.5 Characteristics of the Mobile Management Solution

The mobility management solution proposed in this paper has some advantages. It is easy to deploy because the host's protocol stack does not need to be modified, and no new network entities are needed. It is robust because the mapping information is organized using DHT, and it is secure

because the host cannot access the mapping system directly.

The proposed solution has the following four characteristics:

(1) shortened distance between the router and the mapping system. This is because the mapping systems are deployed in different RAs in a distributed manner; and as a consequence, handover delay is decreased

(2) support for unicast, multicast, and broadcast without any modification of the multicasting protocols. The OA router records the IDs and locators of the hosts that join a multicast group. When it receives the multicast data, it delivers the data to the related hosts

(3) support for multi–interface hosts. Thus, multihoming schemes can be deployed. Each interface of a host can obtain a unique local locator from the local OA router. As a result, the host can receive data via multiple interfaces. The OA router maintains a connection table for the host, which contains the host's ID, the CH's ID, and the local locator of the host's interface used to receive the data. When the OA router receives data destined for the host, it looks up the local locator in the connection table and forwards the data to the corresponding interface. Typically, a host can select the optimal interface based on factors such as access network status, user preference, network cost, and application type. It then updates the local locator saved in the OA router

(4) the address space used by the host is not included in the routing table of the core network. So no routing scalability problem will be caused.

3 Deployment of the Mobility Management Solution

Deployment of the suggested mobility management solution involves CA and RA functions. For instance, the CA and RAs can be deployed in an autonomous area. The functions of CA can be deployed on the backbone router of the autonomous area without any modification. The autonomous area can be divided into several RAs by geographical location. For example, a province can be an RA, and a city within a province can be an OA. On the egress routers of the RA and OA, the functions of RA router and OA router can be deployed. DHT protocol is deployed on these routers to form the mapping system.

4 Conclusion

This paper presents a new mobility management solution based on ID/Locator separation. This solution uses destination address overwriting. It is compatible with the Internet and does not require any new entity to be deployed. The host cannot access the mapping system directly, so security for both users and the Internet is guaranteed. The mapping systems are distributed in different edge networks, and the distance for routers to access the mapping systems is shortened. Therefore, handover delay is also decreased. As a result, users have a better mobile experience. Multicast and broadcast are supported intrinsically as well as multi–interface hosts. Overall theoretical analysis and simulation tests of the system's performance are our near future work.

References

[1] IETF RFC 5944 (2010, Nov.). IP Mobility Support for IPv4, Revised [Online]. Available: http://tools.ietf.org/html//rfc5944
[2] IETF RFC 3775 (2004, Jun.). Mobility Support in IPv6 [Online]. Available: http://www.ietf.org/rfc/rfc3775.txt
[3] IETF RFC 5213 (2008, Aug.). Proxy Mobile IPv6 [Online]. Available: http://tools.ietf.org/html/rfc5213
[4] IETF RFC 4984 (2007, Sept.). Report from the IAB Workshop on Routing and Addressing [Online]. Available: http://tools.ietf.org/html/rfc4984
[5] "BGP routing table analysis reports." [Online]. Available: http://bgp.potaroo.net/
[6] G. Huston, "The BGP Instability Report." [Online]. Available: http://bgpupdates.potaroo.net/instability/bgpupd.html
[7] IETF RFC 5201 (2008, Apr.). Host Identity Protocol [Online]. Available: http://tools.ietf.org/html//rfc5201
[8] IETF RFC 5533 (2009, Jun.). Shim6: Level 3 Multihoming Shim Protocol for IPv6 [Online]. Available: http://tools.ietf.org/html/rfc5533
[9] IETF (2010, Oct.). Locator/ID Separation Protocol (LISP) draft–ietf–lisp–09 [Online]. Available: http://tools.ietf.org/html/draft–ietf–lisp–09
[10] IETF (2009, Oct.). Six/One: A Solution for Routing and Addressing in IPv6 draft–vogt–rrg–six–one–02 [Online]. Available: http://tools.ietf.org/html/draft–vogt–rrg–six–one–02
[11] M. Menth, M. Hartmann, and D. Klein, "Global locator, local locator, and identifier split (GLI–split)," Inst. Comput. Sci., Uni. of Wurzburg, Germany, Tech. Rep. 470, Apr. 2010.
[12] IETF RFC 5204 (2008, Apr.). Host Identity Protocol (HIP) Rendezvous Extension [Online]. Available: http://tools.ietf.org/html/rfc5204
[13] IETF (2010, Oct.). LISP Mobile Node draft–meyer–lisp–mn–04 [Online]. Available: http://tools.ietf.org/html/draft–meyer–lisp–mn–04

Biographies

Yuhong Li (hoyli@bupt.edu.cn) is an associate professor at the State Key Laboratory of Networking and Switching Technology of Beijing University of Posts & Telecommunications, China. Her research interests include mobility management, and next–generation Internet architecture.

Yunjing Hou (houyunjing@catt.cn) is a researcher at Datang Wireless Mobile Innovation Center, China Academy of Telecommunication Technology. Her research direction is mobility management and mobile network architecture.

Shiduan Cheng (chsd@bupt.edu.cn) is a professor and doctoral advisor at the State Key Laboratory of Networking and Switching Technology of Beijing University of Posts & Telecommunications, China. She has long been engaged in the research of telecommunication networks and computer networks.

Self–Adaptive QoS Control in Cognitive Networks That Is Based on Service Awareness

Abstract: This paper analyzes a self–adaptive Quality of Service (QoS) control architecture for cognitive networks (CNs) that is based on intelligent service awareness. In this architecture, packets can be identified and classified using an intelligent service–aware classification model. Drawing on Control Theory, network traffic can be controlled with a self–adaptive QoS control mechanism that has side–road collaboration. In this architecture, perception, analysis, correlation, feedback, decision making, allocation, and implementation QoS mechanisms are created automatically. These mechanisms can adjust resource allocation, adapt to a changeable network environment, optimize end–to–end performance of the network, and ensure QoS.

Keywords: cognitive network; service–awareness; self–adaptive control; QoS

Chengjie Gu
Shunyi Zhang
Yanfei Sun
(Institute of Information Network Technology, Nanjing University of Posts and Telecommunications)

1 Concept of the Cognitive Network

Network service types are varied, and network environments are complex and dynamic. Traditional end–to–end insurance technology lacks intelligent inference and self–learning capabilities. Therefore, it cannot adapt to provide ideal service under dynamically changing network conditions [1].

The obvious problem is that the network system cannot perceive the service demands of end users and cannot effectively and dynamically change QoS according to variations in the internal and external environment of the network system [2]. Academics have started to integrate cognitive elements from next generation networks (NGNs) into current networks in order to overcome these embedded defects. Consequently, the concept of cognitive networks (CNs) has arisen.

Research on CN is focused on the cognitive radio (CR). Mitola [3] first put forward CR and the architecture of the cognitive ring. A CR system obtains frequency spectrum through perception. It determines the reconstruction scheme of CR according to the optimization object and can adapt to changes in the frequency spectrum environment.

CNs based on CR were conceived by the Motorola and Virginia Tech companies [4]. A CN has cognitive processes and perceives the current network condition. It perceives changes in itself and in the environment. It then makes plans and determinations and takes action based on these perceptions. The FOCALE architecture of dual close–loop control is also provided.

At SIGCOMM 2003, Clark [5] and others proposed introducing Knowledge Plane (KP) to the Internet. The key to this concept is that KP can perceive its own behavior. It can analyze problems and adjust its operation to increase reliability and robustness.

In 2007, Baldo [6] used fuzzy logic to process modularization and inaccuracy effectively in the CN. In 2008, Siebert [7] pointed out that the ability of a CN to implement tasks through autonomous self–management, self–optimization, self– monitoring, self–maintenance, self–protection, and self–healing was an important feature. In 2009, Fortuna [8] suggested that Thomas's definition of CN was incomplete. Knowledge expression and cognitive ring are the most important elements of the CN.

The IEEE is currently discussing standardization of the integration architecture of isomerism wireless access networks. In these discussions, the concept of CN is used. CN is seen as a new way of improving overall network and end–to–end system performance as well as simplifying network management. It is the trend of

This work was funded by the National High Technology Research and Development Planning ("863" Project) under Grant No. 2006AA01Z232, 2009AA01Z212, 2009AA01Z202, and the National Natural Science Foundation Project under Grant No. 61003237.

next–generation communication [9].

CN is a new research area and has just taken its first steps in China and in other countries. Therefore, relevant theories and techniques need to be further studied.

The cognitive functions of a CN are implemented by distributed intelligent agents based on AI technology. Agents with learning and reasoning capabilities are deployed on each node in the network to monitor and collect environment information. These agents cooperate and exchange information so that the network can perceive its current status. End–to–end targets can be achieved based on the network status, and network resources can be evaluated, predicted, planned, adjusted, and allocated based on a knowledge library. As a result, the network has self–perception, self–learning, self–optimization, self–healing, and self–configuration capabilities. It can be measured, controlled, managed, and trusted.

In the QoS control architecture of the CN in this paper, the key concept is the network's ability to perceive changes in the CN environment and adjust itself in real time.

Self–adaptive control technology can plan and allocate limited network bandwidth effectively so that network performance is improved. The technology also manages and controls network traffic according to service features in order to improve the revenue of unit bandwidth. Therefore, intensive self–adaptive control is essential to solving network QoS problems in CNs.

2 Key Technologies of Cognitive Network QoS

According to the definition given by Thomas [4], the structure of a CN is described as "Target–Cognitive Decision–Reconfiguration," as shown in Fig.1. The target layer reflects the target demands put forward by the application, user, or resource. With Cognitive Specification Language (CSL), the targets are mapped to specific mechanical demands and fed back to one or more relevant CN elements. The cognitive decision–making layer implements status switching of NEs and perceives the current network status according to the requirements of the target layer. It obtains the NE configuration through certain methods.

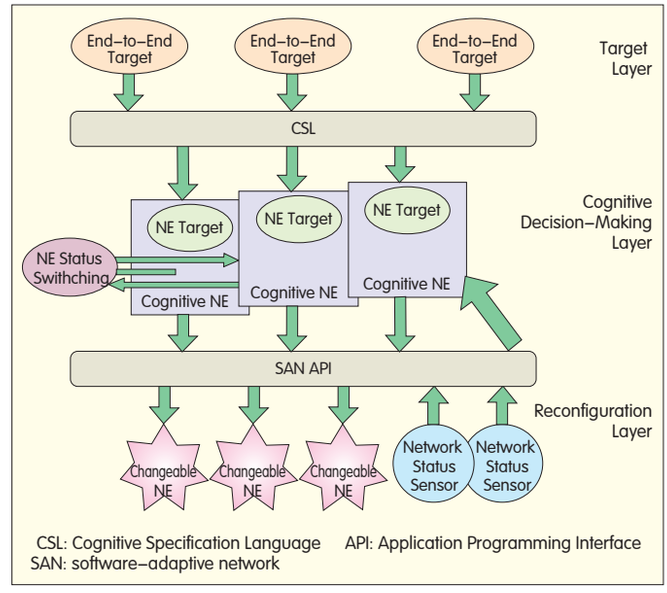


Figure 1. Layers of the cognitive network.

The reconfiguration layer is also called the adaptive network layer. The decision of the cognitive decision–making layer is sent to the corresponding entity NE through the Application Programming Interface (API). By adjusting the configuration of the NE, demands of the target layer are met. At the same time, the layer sends the network status through a sensor to the cognitive decision–making layer.

2.1 Context Perceiving

The foundation of CN is the rapid perception of network environment. A CN needs to observe current network environment information in appropriate time. The information is used in later planning and decision making to determine whether the current network meets user requirements. If not, a suitable reconfiguration method is used to meet user requirements.

Environment information perceived by the CN includes network type, network topology, available resources, interface protocols, and network traffic, all of which affect end–to–end transmission performance [10]. Context perception is an important way of improving network intelligence. It determines changes in context information and adjusts itself accordingly. When the network environment changes dynamically, the network makes relevant self–adjustments. This self– adjustment uses a reflection mechanism and a policy mechanism. From the policy definition, the network can pre–define an adjustment method when the context changes.

2.2 Cross–Layer Design

The essence of cross–layer design is to break the frame of the traditional network system to meet QoS requirements of the communication system. In this design, the status parameters and QoS parameters of the communication system resources are transmitted in the protocol layer. As a result, a joint design combining various protocol layers is achieved, and system resources are fully utilized in order to provide better service for users [11].

The purpose of CN is to adjust the relevant NE protocol stack or protocol layer parameters on the basis of CN network information. This ensures users receive high quality end–to–end performance. The cognitive processing layer knows the status of network layers and determines proper actions according to an optimization algorithm. It reconfigures network parameters and

protocol stacks to achieve end-to-end communication.

2.3 Reconfiguration

If the network is not meeting the end-to-end requirements of users, the CN adjusts the protocol stack parameters of the relevant NE to meet these requirements. The adjustment process is the reconfiguration of the network [12]. CN emphasizes the end-to-end target, and it should provide end-to-end reconfigurability. Software radio technology is limited to reconfiguring the terminal, but CN involves all layers of the NEs and protocol standards that a stream passes through. It is a scheme with foresight that ensures QoS targets are met. More factors are considered in end-to-end reconfiguration.

Realization of CN is based on reconfiguration of the NE. The reconfiguration process can also be implemented through software, but the technical level of this reconfiguration is higher. Terminal reconfiguration, network reconfiguration, and service reconfiguration are contained, and this configuration is not limited to a single node. Multiple NEs on the end-to-end path are covered. This is called end-to-end reconfiguration (E2R). The complexity and importance of E2R is greater than terminal reconfiguration.

3 CN QoS Control Architecture Based on Service Awareness

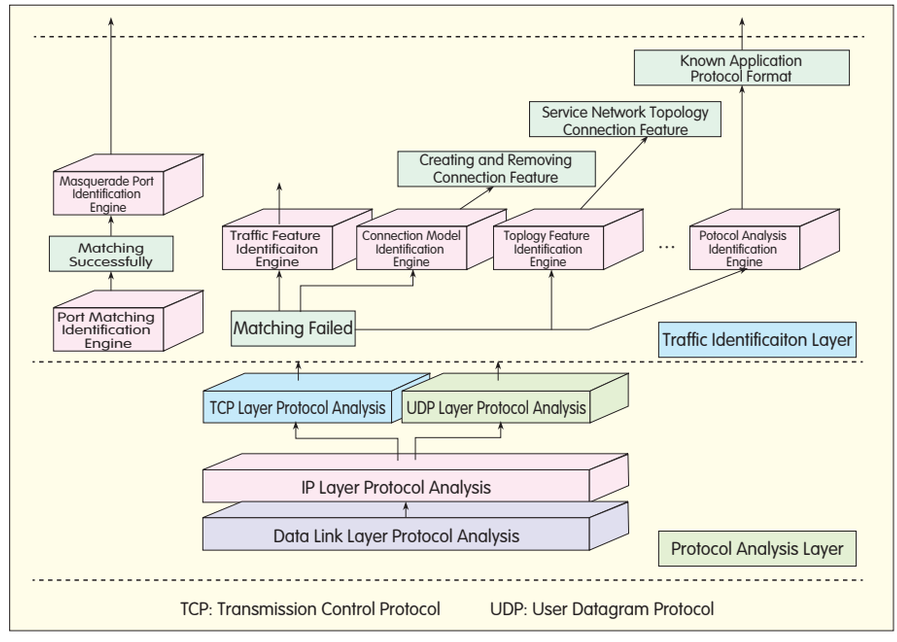
3.1 Service Awareness

In recent years, many new applications have emerged, including peer to peer (P2P) networks, VoIP, streaming media, interactive online games, and virtual reality. The emergence of these new services is impacting the traffic model and application mode. The rapid development of P2P in particular has caused explosive growth in traffic, and unlimited bandwidth usage has increased the burden on the network.

As a result, network congestion has become more serious. Simple expansion cannot meet the requirements of increasing services. Therefore, the best way to perceive, analyze, determine, and control transmission service intuitively is by using CN technology.

CNs are driven by services. The network system intuitively perceives services on the network, including end user service status and NE service status. Intuitive perception and classification based on the service stream is the foundation of service-centered resource configuration, route adjustment, and dynamic self-adaptive traffic control. In service-aware technology, before the CN is introduced, traditional static port method, payload feature method, and stream statistical feature method are used. The methods are effective for perceiving regular services, but they cannot perceive many new services accurately.

After CN is introduced, the network has intelligence as well as analysis and decision-making capabilities. In this section, an integrated feature-based service awareness model is constructed for perceiving services intuitively and intelligently in real time. As shown in Fig.2, after the model has obtained regular parameters of the CN, it constructs an integrated feature identification model based on the traffic statistics feature, connection mode, topology feature, and content feature. It constructs an identification engine for each feature and triggers and perceives different identification engines intelligently according to policies.

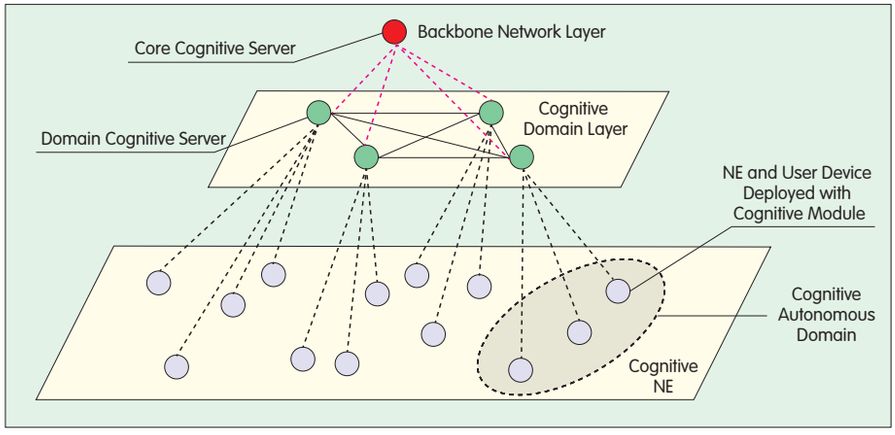


▲Figure 2. Intelligent service awareness and classification model based on integrated features.

As a result, known or unknown, encrypted or plain text traffic can be identified accurately and efficiently. The integrated feature-based intelligent cognitive model distinguishes known or unknown, encrypted or plain text services. This forms the technical basis for a CN self-adaptive QoS control architecture based on service awareness.

3.2 Three-Level QoS Control Architecture in a CN

As shown in Fig.3, the network QoS decision-making and control architecture has a three-level structure, composed of NE (device) cognitive module, autonomous domain cognitive server, and central cognitive server. Each part provides cognitive capability (self-awareness, self-learning, and self-decision making). The NE (device) cognitive module is the basic unit of the CN QoS awareness, analysis, and control system. It provides awareness



▲Figure 3. Three–level self–adaptive control architecture of CN QoS.

and decision–making capability and dynamically adjusts NE parameters or configuration. The NE and user–end devices deployed with cognitive module form a cognitive autonomous domain (configured with a domain cognitive server) that is responsible for managing and controlling the NE device, service traffic, and network resources.

At the same time, a central cognitive sever configured in the architecture is responsible for monitoring, awareness, and management of the running status of the entire network. The layered structure reduces the load on the central cognitive server. Even if the server fails temporarily, service QoS guarantee and management throughout the entire network is not affected.

Distributed networking and communication is enabled between autonomous cognitive servers so that information is exchanged in real time. The reason for using distributed management in the domain cognitive servers is to increase system reliability, flexibility, and expansibility. In the autonomous domain, adjacent nodes communicate so that distributed cooperative monitoring and self–adaptive processing is possible. The architecture integrates the features of centralized architecture and distributed processing technology.

3.3 Port and Path Collaboration in Self–Adaptive QoS Control

CN end–to–end QoS is guaranteed by cognitive NEs. Cognitive NEs are cooperative or independent. The NEs perceive the network condition in real time, bring the trends together, and analyze the network condition. They configure themselves based on existing policies for achieving end–to–end QoS targets.

The following describes the integration of service source–end QoS control and link QoS control in the CN based on service awareness, resource appointment concept, and control theory. A collaborative port and path policy–based self–adaptive QoS control mechanism is proposed to solve the problem of end–to–end QoS guarantee for service traffic. The mechanism sends real–time network parameters to the autonomous domain server (or central cognitive server) through a feedback control.

As a result, the self–adaptive QoS control mode is integrated into the terminal NE and routers. The history of the network condition is compared with the current condition to form a control policy and to update the policy library through self–learning. At this point, the control policy is optimal. The mechanism can ensure the normal operation of a single NE and has the features of CN. The mechanism uses relevant NE devices and reasonably allocates limited resources to improve end–to–end QoE and QoS. In this way, the performance of the entire network is optimized. Fig. 4 shows the awareness–based service source end control layer and distributed awareness–based link control layer.

Awareness–based service control at the source end is implemented through self–adjustment of the source–end transmission rate, intuitive closing of service, and intuitive decrease of QoS target. When the service source end launches a service in the traditional network, the current network condition is not considered. The CN service end has certain cognitive functions; therefore, the cognitive information comes from the domain server or central control server.

When certain conditions are met—for example, when bandwidth is sufficient—resources in the network can accept the access of other service traffic, and the service traffic can be transmitted to the peer end. When high–priority users need to transmit services, but the current network does not provide sufficient resources as required by the SLA, the central cognitive server (or domain cognitive

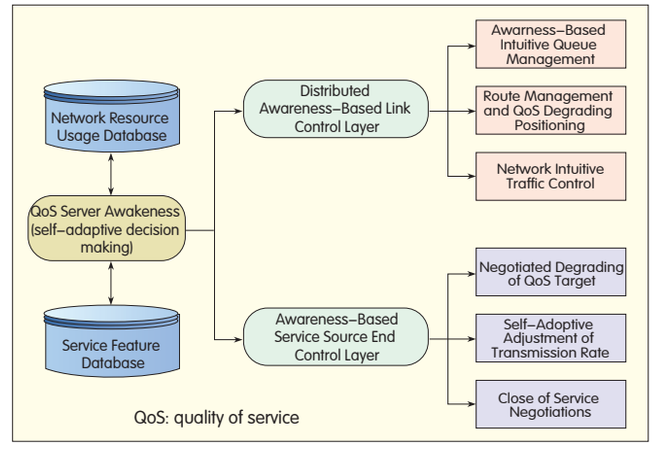


Figure 4. ▶ CN QoS control frame based on port and path collaboration.

server) negotiates with users at the service source end. If the user accepts a reduction of QoS, the source end transmits the service traffic according to the negotiated results. If QoS requirements cannot be reduced, the cognitive server recycles the network resources being used according to the resource distribution policy or even forcibly closes certain low–priority services.

Link control based on distributed awareness is implemented through intuitive control of NE traffic, route management, QoS degradation positioning, and intuitive queue management. By perceiving the network and making decisions based on this information, switches or routers with cognitive functions in the network can intuitively control traffic of different services and ensure the volume of trusted service traffic and key service traffic. They can also limit the volume of unsafe traffic or non–key traffic.

As service requirements and network resources are changing in real time, bottlenecks or QoS–degrading parts of the end–to–end network can be detected by cognitive route management and QoS degradation positioning. In addition, analysis and decision making can be performed, and service traffic can be re–routed. An intelligent and intuitive queue management algorithm can also be used to determine congestion in the CN.

Awareness–based intuitive queue management is oriented to the server's collaborative drive policy. This policy is integrated into the intuitive queue management method to improve the resource appointment algorithm and router buffer management mode. Resources of the router or end system can then be reserved.

4 Conclusions

Owing to the complexity, isomerism, and ubiquity of the access mode and network applications, current networks cannot meet the QoS requirements of users. CN is considered a new way of improving entire network performance and end–to–end system performance as well as simplifying network management. It is the trend of next–generation communication. CNs are important for ensuring performance in complex and isomerism networks.

This paper proposes a self–adaptive QoS control architecture for CNs. With service awareness, self–adaptive control can be implemented in a CN. This architecture is a new approach for solving the problem of NGN end–to–end QoS. Some techniques have been applied in the experimental platform of the national project 863 "Key Techniques in Network Behavior Model–Based Cognitive Network QoS." Monitoring devices are also created and applied for optimizing the network of a carrier. The technique demonstrates sophisticated applicability and stability.

References

[1] C.Lin, Zhichang Shan and Fengyuan Ren, *QoS of Computing Networks*, Beijing: Tsinghua Uni. Press, 2004.
[2] C.Lin, Zhichang Shan, Fengyuan Ren, "QoS of next generation networks," *Chinese Journal of Computers*, vol.31, no.9, 2008, pp.1525–1535.
[3] J. Mitola and G. Maguire, "Cognitive radio: Making software radios more personal," *IEEE Personal Commun.*, vol.69, no.8, 1999, pp.13–18.
[4] R. Thomas, Cognitive Networks. Blacksburg, VA: Virginia Polytechnic and State University, 2007.
[5] D. Clark, C. Partrige, J. Ramming et al., "A knowledge plane for the Internet," *in Proc.Conf. on Applications, Tech., Architectures, and Protocols for Comput. Commun.(SIGCOMM'03)*, New York, pp.3–10.
[6] N. Baldo and M. Zorzi, "Fuzzy logic for cross–layer optimization in cognitive radio networks," *IEEE Commun. Mag.*, vol.46, no.4, 2008, pp.64–71.
[7] M. Siebert, "Self–X control in (future) mobile radio networks," in *Proc.European–Chinese Cognitive Radio Syst. Workshop*, Beijing,2008.
[8] C. Fortuna and M. Mohorcic, "Trends in the development of communication networks: cognitive networks," *Computer Networks*, vol.53, no.9, 2009, pp.1354–1376.
[9] F. Shao and Lifeng Wang "Cognitive network structure and approach based on cognitive level," *J. Beijing Univ. Tech.*, vol.35, no.4, 2009, pp.1181–1187.
[10] P. Balamuralidhar and R. Prasad, "A context driven architecture for cognitive nodes," *Wireless Personal Communications*, vol.45, no.1, 2008, pp. 423–434.
[11] V. Srivastava and M. Motani, "Cross–layer design: A survey and the road ahead," *IEEE Commun. Mag.*, vol.43, no.12, 2005, pp. 88–95.
[12] M. Pitchaimani, B. Ewy, J. Evans, "Evaluating Techniques for Network Layer Independence in Cognitive Networks," *Proc.IEEE Int. Conf. on Commun. (ICC'07)*, Glasgow, 2007, pp.6527–6531.

Biographies

Chengjie Gu (jackiee.gu@gmail.com) is a Ph.D candidate at Nanjing University of Posts and Telecommunications. He is currently researching communication networks, IP technologies, distributed network management, and cognitive networks.

Shunyi Zhang (dirzsy@njupt.edu.cn) is a professor at Nanjing University of Posts and Telecommunications, a Ph.D tutor, a member of China Communication Academy and the director of the IP Application and Value–added Telecommunication Technique Commission. He is also the associate director of the China Electronic Academy, and director of Jiangsu Province Engineering Research Center of Telecommunication and Network Technology. He is currently researching the computer network communication, communication network and IP technology.

Yanfei Sun (sunyanfei@njupt.edu.cn) is an associate professor at Nanjing University of Posts and Telecommunications. He is also a master's tutor. He is currently researching the network performance monitoring and optimization, QoS control and management, and multimedia network communication.

Roundup

BT and ZTE Announce Research Partnership

One June 01, 2011, ZTE Corporation announced a research partnership with BT to explore the next generation of fixed line, wireless and mobile telecom services.

The partnership will see ZTE's extensive knowledge of equipment and network solutions combined with BT's world–class innovation capabilities and experience of delivering communications services in more than 170 countries worldwide.

BT and ZTE will look to develop international telecoms standards which drive the interoperability and convergence of global communications systems.

(ZTE Corporation)

A P2PSIP System with Intelligent Routing Function on the Media Plane

Yongsheng Hu[1], *Zhenwu Hao*[1], *Jun Wang*[1], *and Naibao Zhou*[2]
(1. Central R&D Institute, ZTE Corporation;
2.China Mobile Research Institute)

Abstract: Decentralized peer–to–peer session initiation protocol (P2PSIP) provides the same services as legacy SIPs such as IMS. However, in relatively open network, the requirement for route efficiency in a complex environment brings about undefined problems. To deploy a controllable P2PSIP network, perfect mechanisms have to be appended, especially in QoS, security, and management. Several proposals for QoS, network address translation (NAT), and interworking have been put forward. In this paper, we propose an integrated architecture for a P2PSIP system as well as a proactive intelligent routing scheme on the media plane used in system. Implementation and simulation show that our solution is suitable for operation and management.

Keywords: peer–to–peer; session initiation protocol; relay; intelligent routing

1 Introduction

The IETF peer–to–peer session initiation protocol (P2PSIP) working group oversees standardization of decentralized SIP architecture [1]. IETF hopes to avoid centralized architecture, and to create a Skype–like system [2] but with open standards. P2PSIP is only a basic end–to–end communication protocol; building an integrated network suitable for operation and management still requires mechanisms for QoS, security and management, and accounting and billing.

Compared with legacy SIP systems [3] such as IMS, P2PSIP has a more open environment. For example, service nodes are distributed, rather than centralized in a closed house. Peer nodes are usually located behind network address translations (NATs). The network environment is more like the Internet than traditional communication networks with reliable QoS. Therefore, a P2PSIP system requires intelligent routing to solve NAT and route efficiency. In this paper, we propose a novel and practical P2PSIP architecture that uses overlay techniques for signaling and media transport. It replaces the traditional SIP proxy/register function with distributed hash table (DHT) mechanism, and a relay service is added to resolve NAT and QoS problems.

This work was funded by the Next Generation Bandwidth Wireless Mobile Communication Network Program, a Key National Science and Technology Specific Project sponsored by MIIT of China, under Grant No. 2010ZX03004–001.

2 Related Works

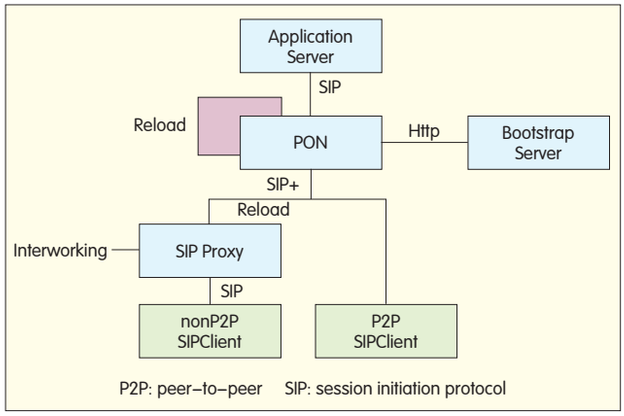
The IETF P2PSIP working group is the first international standards organization focused on evolving SIP towards being more peer–to–peer, that is, removing the centralized proxies and adopting a DHT approach. This working group has proposed a basic resource location and discovery protocol called RELOAD [4] and defines a simple SIP usage [5] based on it. SOSIMPLE [6] and SIPeerior [7] are important projects based on RELOAD.

There are also several voice over IP (VoIP) systems that are non–SIP or non peer–to–peer. These include Skype [8], Gizmo [9], and SATO [10]. Skype can be seen as the driver of some peer–to–peer SIP developments because it has shown how well VoIP can work and what a successfully deployed VoIP peer–to–peer system looks like. Compared with these systems, our P2PSIP focuses more on control, operation, and management.

Intelligent routing is used to inspect the status of the network and select an alternative route that excels the default Internet routing. Intelligent routing is realized by application–level overlay routing, that is, relay technique. The relay function is key for a P2P VoIP system to guarantee NAT and QoS. The main advantage of Skype is that it uses the equivalent of session traversal utilities for NAT (STUN) servers and traversal using relay NAT (TURN) servers in the node itself when handling NAT [2]. This is unlike explicit server pre–configuration in existing SIP applications. Based on Internet routing behavior analyzed by Chinoy [11] and Labovitz [12], S. Savage et al [13] proved that most default/bad IP paths could be improved by alternative paths in the Internet. In the last decade, many studies have focused on overlay routing such as RON [14], Detour [15], SOSR [16], and ASAP [17].

3 Design of P2PSIP architecture

Our goal is to develop a flexible and controllable P2PSIP network for



◀Figure 1. The proposed P2PSIP system architecture.

traditional telecom operators. This is different from Internet P2P telephone Skype in terms of QoS, service control, and network management.

3.1 Design Rules

We deploy a decentralized SIP system in a virtual peer–to–peer network. The requirements are:

• overlay self–organization. The P2PSIP network consists of P2PSIP overlay nodes (PONs) on top of the underlying basic network connectivity. This virtual overlay network can be self–tuned while network adjusting and disturbing occurs. In the future, it will run in a virtual machine environment.

• generic SIP service support. Existing basic and important IMS functions, such as session control, authentication, roaming, and interworking, should be supported by P2PSIP. Key techniques, such as data modeling and creating database of user service profile, as well as triggering of SIP service, must be solved.

• terminate node access. In a decentralized environment, user authentication and access security are undefined problems.

• interworking. As an alternative to VoIP, there must be a way to route calls between the P2PSIP network and traditional SIP network and PSTN. To solve this issue, gateways act as either a P2P or non–P2P client.

• intelligent routing function. The main purposes of relay service are NAT traversal and QoS proxy. Potential functions of relay nodes include audio/video mixer, audio/video transcoder, and interceptor.

The current developed P2PSIP system does not have all these functions. But we provide several useful mechanisms in our design. For example, by anchoring a user's service on a specified PON, all services are processed at a fixed point. This allows convenient service control and management.

3.2 The P2PSIP System Architecture

There are different components in the proposed P2PSIP system (Fig. 1). The core of the P2PSIP system is the PON. It processes SIP services and cooperates with other PONs to implement the P2P overlay network. The protocol between PONs is RELOAD [4] defined by IETF. RELOAD is currently the only available standardized P2P protocol.

The bootstrap server is the first node when a PON joins the P2PSIP overlay network. The overlay configuration file is provided by the bootstrap server. The selected protocol between the bootstrap server and PONs is HTTP. The overlay configuration file is encoded and transferred in XML format.

The application server hosts and executes IMS–specific services such as the voice call continuity (VCC) functions and presence–based services. This element–and its interface with the P2PSIP overlay network–inherits aspects of legacy IMS components.

Both P2P and non–P2P clients are allowed to access the P2PSIP system. The SIP proxy provides protocol transfer for the non–P2P client and controls interworking with other telecom systems, , for example, connecting with breakout gateway control function (BGCF) and media gateway control function (MGCF).

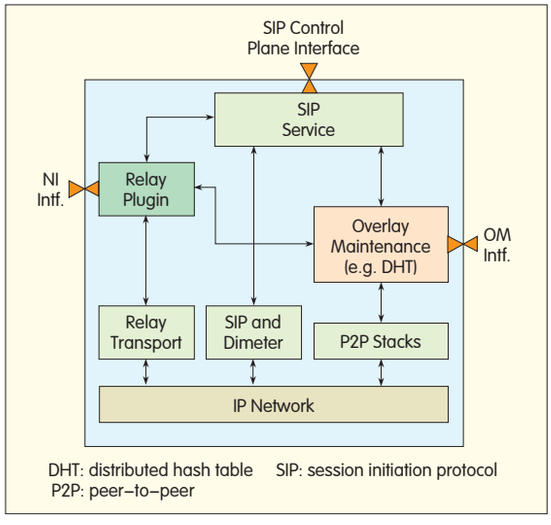
3.3 Intelligent Routing Function

NAT is a basic feature of legacy IMS networks in real network environments. The media relay mechanism is necessary for improving best–effort route policy in the Internet. Therefore, we propose a relay service based on P2P overlay function in PONs.

A PON supports the SIP service function, relay service function, and overlay maintenance function. Acting as a supporting layer, the overlay maintenance function organizes all PONs into an integral overlay network according to the specific DHT algorithm. The P2PSIP system is an application–level service overlay network established on the low–level P2P overlay system. The intelligent routing function in the P2P overlay network maintains a candidate relay–node list that provides preferred relay–nodes. Fig. 2 shows the PON architecture.

We propose a smart relay function based on the supportive P2P overlay network (Fig. 2). The relay plugin maintains a topology–aware candidate relay table. This table is based on the P2P overlay and global network information provided by a centralized server similar to the ALTO server defined by IETF [18]. This relay table is created in two steps when a PON joins the P2P overlay. First, the new PON obtains its local network information (including network partition and neighbor network partition) from the network information server via the NI interface. The individual network partition is identified by a packet identifier (PID) and may be one or more autonomous systems (AS) and one or more subnets. Then, the new PON collects candidate relay nodes belonging to itself and neighbor PIDs using its DHT routing table. During overlay maintenance, the relay table is also updated. A relay table structure is shown in Fig. 3.

A relay node should meet following four requirements [13], [14], [17]:



▲Figure 2. The PON module design.

(1) quasi real–time path QoS inspection. If the path QoS deteriorates badly, a relay node should apperceive and update its relay table as fast as possible. The relay plugin is required to measure the quality of the path to its neighbor nodes.

(2) node workload and capability perception. To select a suitable relay node, node workload, capability, and volunteer status needs to be taken into account. The relay function should apperceive the state of neighbor nodes in its relay table.

(3) shortest path first for data relay. Previous research [16], [17] has proven that a one–hop routing path can satisfy the requirements of voice communication. The proposed distributed scheme based on DHT overlay is capable of quick one–hop path selection in AS–level.

(4) closest node first for NAT traversal. In a real Internet environment, it is essential to provide proxy node selection for consumers behind NAT devices. Our system dynamically provides a closest relay node for NAT as opposed to manual pre–configuration in the traditional mode.

4 Implementation of Proposed P2PSIP System

Our P2PSIP system is implemented on the premise that traditional SIP user equipment (UE) can participate without any enhancement. The SIP UE uses its standard way of transporting SIP signals and real–time transport protocol (RTP) media. This is similar to IMS terminating with a preconfigured access PON node. In the future, advanced P2PSIP UE will select an access PON node automatically using the P2P client described in [4].

A call made between UE_1 in PID_1 and UE_2 in PID_2 follows the possible call flows in [3]. The preconfigured access PON node (PON_1) for UE_1 receives an INVITE call request, obtains the UE_1 user profile (by RELOAD Fetch request), and triggers its subscribed services. Then, PON_1 uses SIP URI to obtain UE_2 register information (by DHT lookup) and helps build up a communication channel between UE_1 and UE_2 [3].

4.1 Relay in Media Path

If end–to–end QoS between UE_1 and UE_2 gets worse during a session, an UPDATE request is sent to PON_1 (the current session process node). After PON_1 checks the UE_1 user profile and finds the user subscribed to the relay service, a relay selecting flow is triggered.

The relay request is not necessarily routed to PON_1 but to a PON that is closest to UE_1. This routing is performed by the RELOAD mechanism (Fig. 4). The peer is a surrogate node (SN1) that selects a relay node for the session. The SN_1 is located at PID_1 or close to PID_1 if no peer exists in PID_1. The SN_1 finds a surrogate node for UE_2 (by its location, PID_2) called SN_2, and obtains its relay table (Fig. 3). Based on the four defined principles for relay selection, the SN_1 selects one or more relay candidates by ranking.

4.2 NAT Traversal Process

When PON_1 receives the INVITE request and finds UE_1 or UE_2 behind a NAT device, a relay is required for data packet transfer. A NAT relay request for UE_1 in PID_1 is routed to a PON that belongs to PID_1 or is closest to PID_1 (Fig. 5, PON_1). The PON_1 selects the closest neighbor node with a light load in the local network partition (the first row in its relay table). Here, the metric of distance is the most prefix match of IP addresses.

4.3 Environment and Implementation

The designed P2PSIP system has been implemented on Open SUSE 11.2 in C++ language . On the network side, an open source stack reSIProcate [19] was used. P2P stack is a self developed project based on IETF RELOAD [4]. Test SIP phones include the well–known Eyebeam SIP phone and SoftDA (ZTE testbed). Other devices, such as application server, NAT devices, and IMS system for inter–communication test, are all ZTE commercial products.

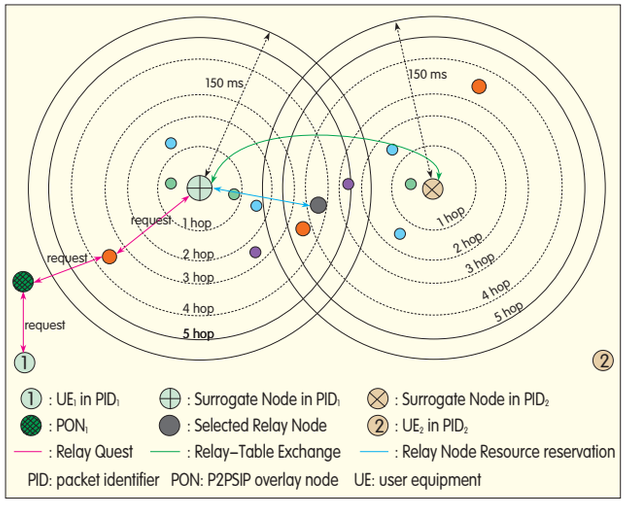
5 Conclusions

We propose a novel P2PSIP system with intelligent routing function in the media plane. The system follows SIP

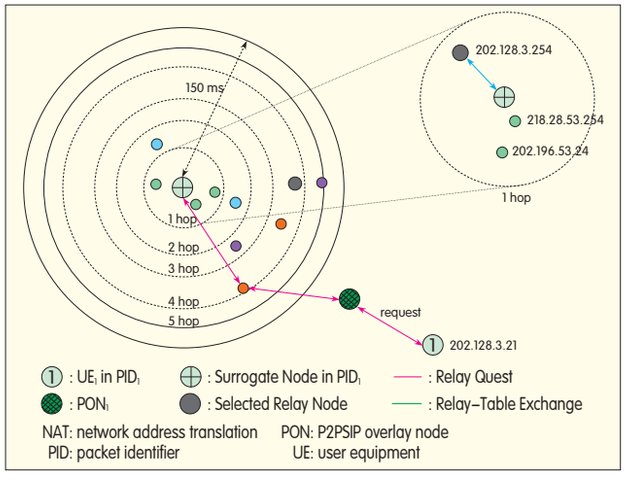
No.	Hops	PID	Neighbor Node List
1	0	PID_0	List<T NEIGHBOR NODE>
2	1	PID_2	List<T NEIGHBOR NODE>
3	1	PID_4	List<T NEIGHBOR NODE>
4	1	PID_8	List<T NEIGHBOR NODE>
5	2	PID_{23}	List<T NEIGHBOR NODE>
6	2	PID_9	List<T NEIGHBOR NODE>
7	2	PID_{13}	List<T NEIGHBOR NODE>
8	3	PID_{21}	List<T NEIGHBOR NODE>
9	3	PID_{97}	List<T NEIGHBOR NODE>
…	…	…	…

```
typedef struct {
  NodeID       node_id;    // PON  nodeID
  IPAddrPort   node_addr;  // PON node info, IP and PORT
  T_NODE_CAP   node_cap;   // PON node capability,
                           // CPU, Mem, net, etc.
  T_NODE_LOAD  node_load;  // PON node workload
  T_QoS        path_qos;   // path QoS to local
                           // RTT, loss, jitter, etc.
  …
} T_NEIGHBOR_NODE;
```

▲Figure 3. An Example of relay table structure.



◀Figure 4.
P2P calling setup process.



◀Figure 5.
NAT traversal process.

functions and features of 3GPP IMS services. Basic VoIP services are supported; for example, an SIP session can be set up between traditional UEs. For UE behind NAT devices, a specified application server selects an appropriate relay for users. If the default network path gets congested, a relay service is provided as the subscribed service.

A key advantage of P2P technique is that distributed process and storage eliminates the centralized point failure. However, during implementation of our P2PSIP system, we found that using DHTs affects performance in the more flexible routing path in the application layer. Furthermore, the decentralized architecture makes deployment of the traditional SIP/IMS services, such as multiple user identifiers (PUIs and PVIs) and forking service supporting, more complicated. Security in the overlay network and control of services are also major concerns. These issues are still to be investigated.

References

[1] "P2PSIP: Peer–to–Peer Session Initiation Protocol." [Online]. Available: https://datatracker.ietf.org/wg/p2psip/
[2] S. A. Baset and H. Schulzrinne, "An analysis of the Skype peer–to–peer Internet telephony protocol," Cornell University, Ithaca, NY, Rep. CUCS–039–04, 2004.
[3] *SIP: Session Initiation Protocol*, RFC3261, June 2002.
[4] C. Jennings, B. Lowekamp, E. Rescorla, S. Baset, and H. Schulzrinne, "Resource location and discovery (RELOAD) base protocol draft–ietf–p2psip–base–13," IETF. [Online]. Available: https://datatracker.ietf.org/doc/draft–ietf–p2psip–base/
[5] C. Jennings, B. Lowekamp, E. Rescorla, S. Baset, and H. Schulzrinne, "A SIP usage for RELOAD draft–ietf–p2psip–sip–05," IETF. [Online]. Available: http://tools.ietf.org/html/draft–ietf–p2psip–sip–05
[6] D. A. Bryan, B. Lowekamp, and C. Jennings, "SOSIMPLE: a serverless, standards–based, P2PSIP communication system," in *1st Int.l Workshop, Adv. Architectures and Algorithms for Internet Delivery and Applic.* (AAA–IDEA), Orlando, FL, June 2005, pp. 42––49.
[7] "SIPeerior Technologies P2PSIP Core and Endpoint Development Kits," P2PSIP. org. [Online]. Available: http://www.p2psip.org/implementations.php
[8] Skype. [Online]. Available: http://www.skype.com
[9] Gizmo. [Online]. Available: http://gizmoproject.com
[10] M. Stiemerling and M. Brunner, "A peer–to–peer SIP system based on service aware transport overlays," *PIK*, vol. 30, no. 4, pp. 213–218, 2007.
[11] B. Chinoy, "Dynamics of Internet Routing Information," in *Proc. SIGCOMM'93*, New York, NY, 1993, pp.45–52.
[12] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, "Delayed Internet routing convergence," in *Proc. SIGCOMM'00*, New York, NY, 2000, pp. 175–187.
[13] S. Savage, A. Collins, E. Hoffman, J. Snell, and T. Anderson, "The end–to–end effects of internet path selection," *in Proc. SIGCOMM'99*, New York, NY, 1999, pp. 289–299.
[14] D. Andersen, H. Balakrishnan, F. Kaashoek, and R. Morris, "Resilient overlay networks," *ACM SIGCOMM Comput.* Commun. Rev., vol. 32, no. 1, p. 66, Jan. 2002, pp.xxx–xxx.
[15] S. Savage, T. Anderson, A. Aggarwal, D. Becker, N. Cardwell, A. Collins, E. Hoffman, J. Snell, A. Vahdat, G. Voelker, and John Zahorjan, "Detour: informed Internet routing and transport," *IEEE Micro*, vol. 19, no. 1, pp. 50–59, Jan./Feb. 1999.
[16] K. Gummadi, H. Madhyastha, S. Gribble, H. Levy, and D. Wetherall, "Improving the reliability of Internet paths with one–hop source routing," In *Proc. USENIXOSDI'04*, Berkeley, CA, 2004, p.13.
[17] S. Ren, L. Guo, and X. Zhang, "ASAP: an AS–Aware Peer–Relay Protocol for High Quality VoIP," in *26th IEEE Int. Conf. Dist. Comput. Syst. (ICDCS' 06)*, Lisboa, Portugal, 2006, p. 70.
[18] R. Alimi, R. Penno, and Y. Yang , "ALTO protocol draft–ietf–alto–protocol–06." [Online]. Available: http://tools.ietf.org/html/draft–ietf–alto–protocol–06
[19] reSIProcate. [Online]. Available: http://www.resiprocate.org/Resip_Overview

Biographies

Yongsheng Hu (hu.yongsheng@zte.com.cn) received his Ph.D. degree from Nanjing University of Science and Technology, China, in 2008. He is a senior engineer in the Central R&D Institute, ZTE Corporation. His current research interests include distributed system, content network and cloud computing.

Zhenwu Hao (hao.zhenwu@zte.com.cn received his M.S. degree from Nanjing University of Science and Technology, China, in 1996. He leads core network standardization research in the System Architecture deptartment of Central R&D, ZTE Corporation.

Jun Wang (wang.jun17@zte.com.cn) graduated from Nanjing University of Aeronautics and Astronautics, China and received his M.S. degree in 2006. He is an architect at the System Architecture deptartment of Central R&D, ZTE Corporation. His research interests include core network evolution, distributed systems, and datacenter networking.

Naibao Zhou (zhounaibao@chinamobile.com) graduated from Beijing University of Posts and Telecommunications, China and received his M.S. degree in 2010. He is an engineer at the China Mobile Research Institute. His research interests include core network evolution and distributed systems.

Architecture and Key Technology of Distributed Intelligent Open Systems

Xiaoyu Tong , Yunyong Zhang , and Bingyi Fang
(China Unicom Research Institute)

Abstract: High–speed large–bandwidth networks and growth in rich internet applications has brought unprecedented pressure to bear on telecom operators. Consequently, operators need to play to the advantages of their networks, make good use of their large customer bases, and expand their business resources in service, platform, and interface. Network and customer resources should be integrated in order to create new business ecosystems. This paper describes new threats and challenges facing telecom operators and analyzes how leading operators are handling transformation in terms of operations and business model. A new concept called distributed intelligent open system (DIOS)—a public computing communication network—is proposed. The architecture and key technologies of DIOS is discussed in detail.

Keywords: DIOS; public computing communication network (PCCN); cloud computing

1 Introduction

Since the reshuffle of the Chinese telecom industry in 2008, Chinese telecom operators have been deploying 3G networks, and upgrade of the Internet has sped up. Competition between operators is fiercer than ever. Considerable bandwidth is now being consumed by a diversity of applications, but telecom operators are not obtaining reasonable ROI on these services. Nor are they adding to the value of their enterprises by capitalizing on the high value–added services. The Internet has become a huge industry that integrates services and applications, creates greater enterprise value, and has significant social influence. It is entering an era of marketing channels, content, and applications, and the dominant players in the industry chain are changing. Radio and TV operators are converting their radio and TV networks into digital networks in order to promote tri–network convergence. They have become new competitors in the information communication service sector.

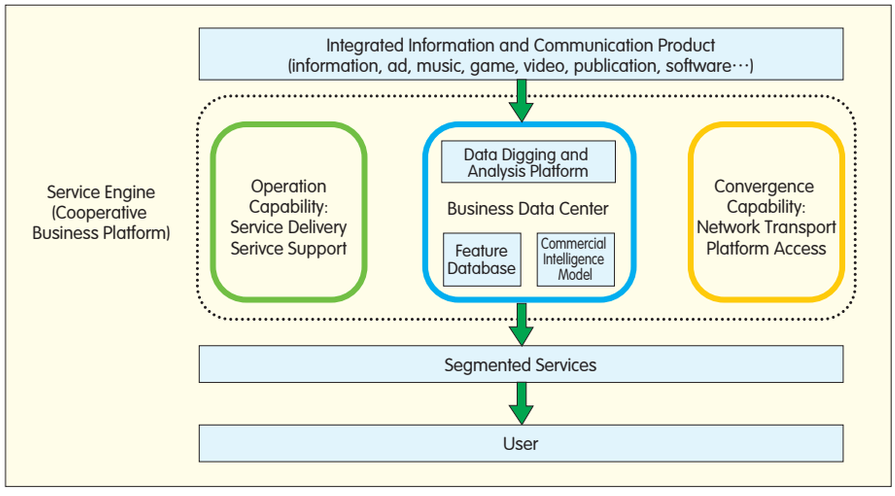
With the arrival of 3G and the development of mobile Internet, IT, the Internet, media, telecom and retail industries are quickening convergence and penetration. Their boundaries are becoming indistinctive, and a new industry ecosystem is coming into being. All participants in this ecosystem need to be open and cooperative so that they can coexist and develop together. A new post–telecom era is coming. Telecom operators are platform providers and should be more open than ever to creating and sharing value, constructing platforms, engaging in standardization, redesigning and optimizing networks, and creating cooperation mechanisms between platforms and terminals.

In light of the current information service environment and the transformation of telecom business models, this paper proposes a distributed intelligent operation system (DIOS). DIOS is regarded as the core system for future public computing communication networks (PCCNs) [1]. DIOS is capable of intelligent network management, intelligent business development, and intelligent service provision. By using and integrating public platform resources, DIOS shortens service innovation cycles and maintains an ecosystem where several industries can co–exist.

2 Objectives of DIOS

DIOS is designed to transform the business models of telecom operators. The transformed models of leading Chinese telecom operators may be one of the following [2]:

- a communication and marketing channel model. Controlled channels, application combinations, and terminal experience are enhanced along with user demand analysis to ensure the operator's products and services satisfy demands. The operator establishes prompt, economic, and on–demand marketing capabilities.
- a communication and media advertising model. The operator exploits rich customer demand information, improves its data analysis capability, and uses its marketing



▲Figure 1.Network architecture after transformation.

channel and terminal delivery capability to develop a "pan–terminal," "rich media," segmented all–round advertising model.

• a communication and information service model. The operator uses the infrastructure and trusted, secure public communication network (PCN)—as well as emerging broadband communication and cloud technologies—to offer services of information facilities, information production, and information application.

• a prepaid and postpaid model. With the channel and media feature–based postpaid model, communication and information expenses of users are reduced. The operator can quicken the growth its user base (quick aggregation), increase service usage time (increased browsing traffic), enhance the channel and media utility, and improve its profit–making capability from agent fees and advertising fees.

In transforming its business model, a telecom operator should make use of its network and customer resources to operate commercial platform and interface resources. To adapt to new communication, marketing channel, media advertisement and information service models, an operator should exploit its convergence capability (for network transport and platform access) and its operational capability (for service delivery and support). It should also add a BDCS core operation system (Fig.1). After the transformation from network–centered operation to customer–centered operation, a communication network is simply a channel for users to access services and a means for user aggregation. All new business models operate as follows: First, the BDCS mines user information. Then, the feature database and commercial intelligence model analyzes the collected information. Finally, customized, segmented services are delivered to users via the PCCN, allowing them to have a better experience on their terminals.

Where network services tend to be information services and information management services, providing services requires a much computing. As network bandwidth increases, computing capability gradually spreads over the entire network. Communication technologies and services are developing towards computing technologies and applications, and computing technologies and applications are developing towards delivery of networks and services. Today's PCN will evolve into PCCN. The DIOS is the core system of future PCCN. It is distributed, intelligent, open, and integrated, and it is capable of intelligent network management, intelligent business development, and intelligent service provision.

Using distributed storage, distributed computing resources, distributed databases, and file systems, DIOS can be used to achieve several objectives. First, it allows for on–demand scalability by integrating all resources into the cloud network. Second, it enables the network to be intelligently self–organized and reconfigurable by means of data mining, analysis, and intelligent scheduling. Third, it finds the best match between services and resources, between applications and services, and between terminals and users. Fourth, it quickly and comprehensively responds to service demands by establishing an open–capability engine platform that allows a third party to access cloud resources. Fifth, it maximizes the sharing of integrated resources by unifying virtualization standards and reducing costs. Sixth, it is capable of carrier–grade operability and manageability because it sets up a unified management platform to improve centralized management efficiency.

3 PCCN

PCCN is an information processing network based on virtualization and cloud computing. It integrates the communication network and computer network. By using cloud virtualization, PCCN establishes a support network, service network, and uniform infrastructure resource pool. By introducing cloud computing, it organizes and uses the infrastructure resource pool efficiently. Using access, switching, routing and transmission elements of existing PCN, the PCCN implements computing processing, virtual allocation, scheduling management, and a service development environment (Fig.2).

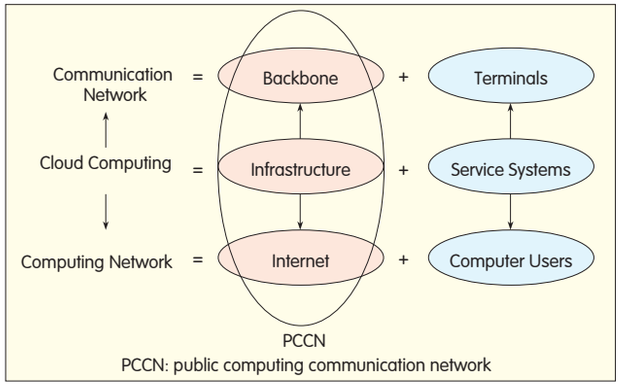
4 DIOS

The DIOS is the core system and the implementation element of PCCN. Fig. 3 shows the six–layer DIOS topology.

In the following section, we will discuss the six layers in detail.

Layer 1: BDCS Layer

The Business data center system (BDCS) layer consists of three



◀Figure2. Relationship between PCCN and communication network, cloud computing, and computer network.

subsystems: network data center, user data center, and operation data center.

Layer 2: Cloud Resource System

The cloud resource system comprises three parts: cloud storage, cloud network, and cloud computing. Cloud storage devices include a massive database and a distributed file system. The cloud network device is a compact frame comprising cloud routers and a programmable, virtual cloud switcher with enhanced resource sharing function. The physical cloud devices include mini computers and X86 servers.

Layer 3: Capability Engine Layer

The capability engine layer also comprises three parts: event processing engine, business control engine, and data analysis engine.

Event processing involves implementing workflow–based business procedures and management procedures such as business support systems (BSS), partner relationship management (PRM), operations support system (OSS), open data services (ODS), and office automation (OA).

Data analysis involves computing structured and non–structured data using specific computing rules to find out general business intelligence (BI), customer relationship management (CRM), and search engine services from mass of the data.

Business control involves strategic control over business, network organization, application adaption, and service delivery according to predefined rules such as session border control (SBC), call session control function (CSCF), subscriber data center (SDC)/home location register (HLR) /home subscriber server (HSS), authentication, authorization, and accounting (AAA), application programming interface (API), and service delivery platform (SDP). Typically, business control involves the business control module of the value–added platform as well as the user credit control module of the billing system.

Layer 4: Cloud Resource Control Node Layer

The cloud resource control node layer is based on distributed architecture and shields the complicated physical and logic structures in the cloud. It uses scalable, adaptive load balancing and dynamic, intelligent resource adaption to match services with service engines. In this way, automatic and intelligent scheduling is achieved.

Layer 5: Cloud Access Gateway Layer

The cloud access gateway layer, comprising physical access gateways and service platform access gateways, allows terminals to access the cloud. It shields the difference between physical devices and service platforms and allows different terminals to intelligently access the cloud. Thus, unified access is achieved.

Layer 6: Cloud Terminal Layer

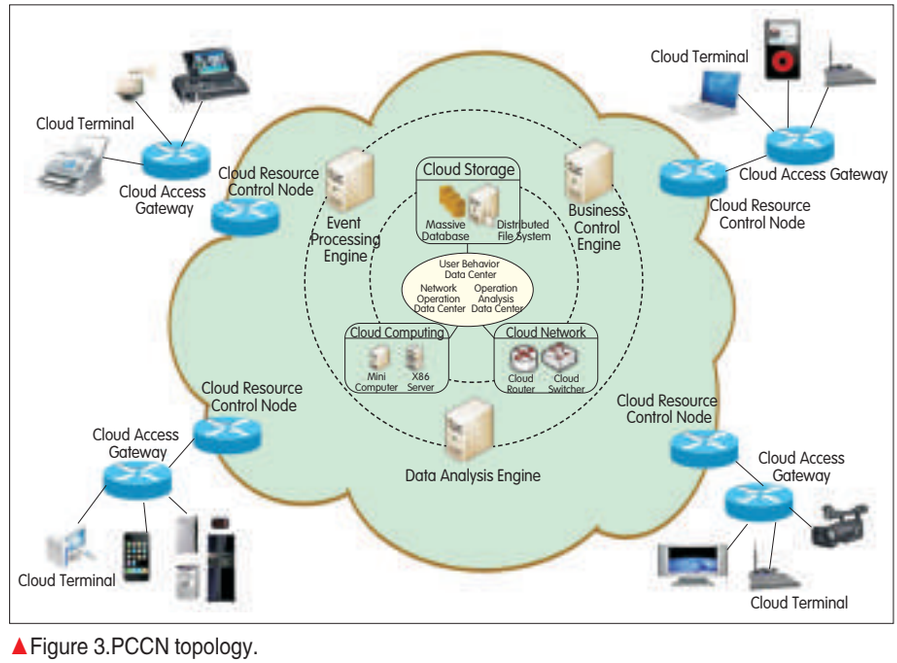
The cloud terminal layer comprises physical devices and client software. The physical cloud devices can be further divided into thin terminals, dumb terminals (such as sensors of the Internet of things), intelligent soft terminals, and browsers.

The DIOS architecture is shown in Fig. 4.

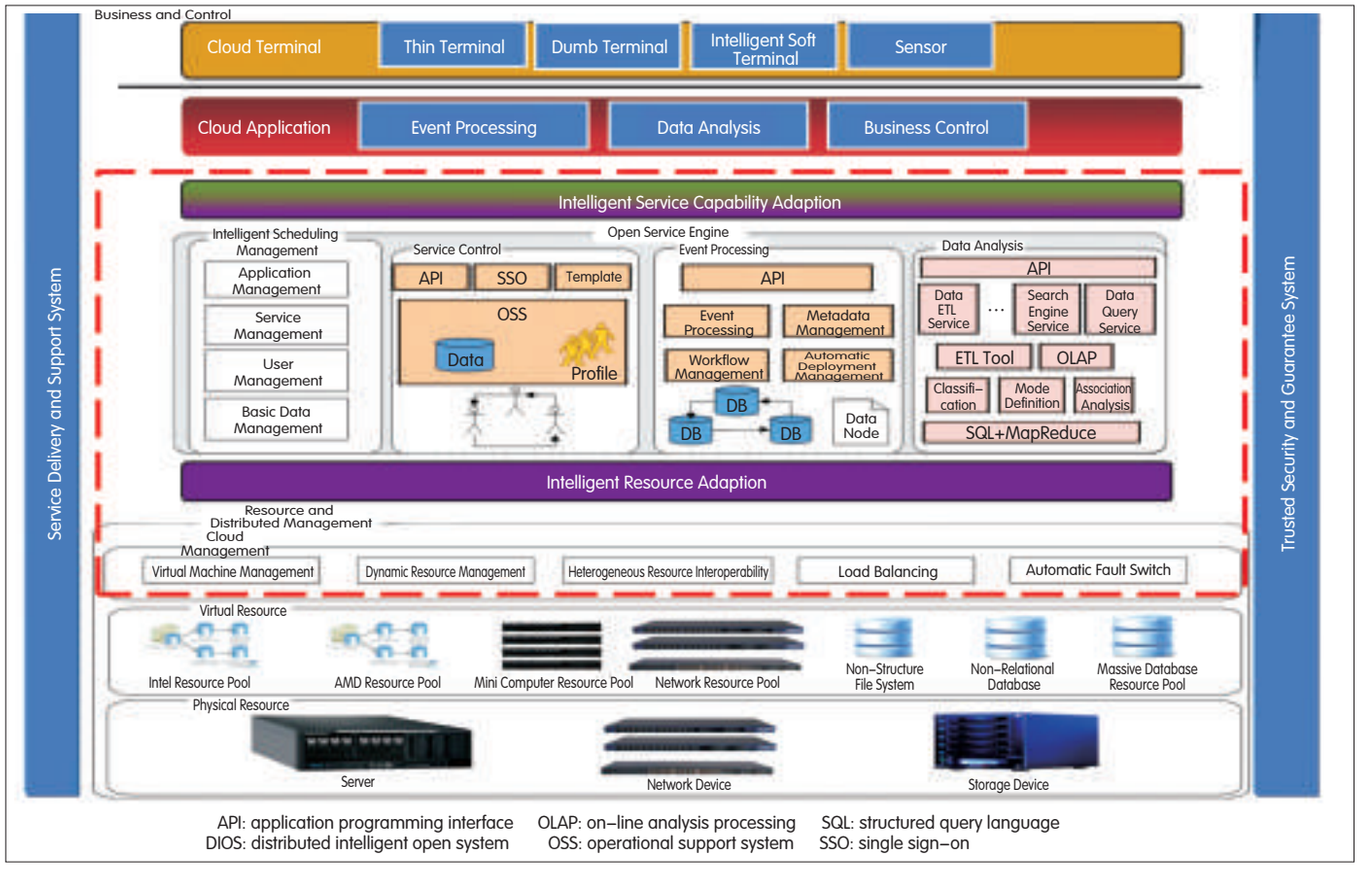
5 Key Technologies of DIOS

5.1 Unified Virtualization

Virtualization technologies are the basis of DIOS, and they fall into several types. Virtualization in DIOS involves server virtualization, storage



▲Figure 3.PCCN topology.



▲Figure 4.DIOS architecture.

virtualization, and network virtualization. The virtualization technologies include heterogeneous resource virtualization, live migration of heterogeneous virtual machine, virtual fault-tolerance and disaster recovery, bearer layer virtualization, control plane virtualization, and reconfigurable intelligent network. Of these, heterogeneous resource virtualization is essential for the coexistence of various systems, and it has become a powerful tool for providing customized resources in the context of network and data center. It provides a logic view rather than a physical view of data, computing capability, storage resources, and other resources. Thus, it shields many physical, structured details. Only with unified virtualization can heterogeneous resources be fully exploited and easily managed.

5.2 Unified Cloud Management

The DIOS matches and schedules resources concurrently for complicated applications in order to maximize the use of resources and enhance execution capability and running performance across the entire network. Thus, users are provided with higher quality services. Unified cloud management includes virtual machine scheduling management (for automatic configuration and scheduling of virtual server resources), deployment management (for automatic installation, configuration, and batch deployment of OS and applications), and storage management (for integrating heterogeneous storage resources, remote synchronous and asynchronous copying functions, remote mirror copy, and snapshot).

5.3 Open Service Capability Engine

Third-party service providers are the main driving force of development in the telecom industry. Creating a diversity of services requires the participation of a large number of third-party developers. Applications that combine different service capabilities and have different characteristics require many third-party users to be part of the development process. Two main functions of a telecom operator are: to create an open service capability engine, allowing third-party developers to develop low-cost services; and to enhance service innovation to attract more users.

DIOS can provide new application scenarios for telecom services. With the open business control, event processing, and data analysis platforms of the capability engine, DIOS opens cloud resources to third-party developers for quick response. The open service capability engine of the DIOS has basic service capability and integrated service capability. The basic service capability includes business control, event processing, and data analysis, and the integrated service capability supports access of various networks to provide ICT services. The

DIOS controls the open service capability engine for uniform user and capability authentication, load balancing, routing allocation, and session control for all service requests from the open service interfaces. The management component of the service capability engine has control functions so that BSS–related and OSS–related functions can be realized. These control functions include access control, strategy management, configuration management, billing management, fault management, monitor management, and statistical analysis.

5.4 Intelligent Resource Adaption

When the service capability engine is open to third parties, resources at the lower layer may be used. The intelligent resource adaption function intelligently matches services with different network and computing resources centrally schedules and manages resources of the service capability engine to maximize resource use in the upper layers.

6 Conclusion

Chinese telecom operators are faced with unprecedented challenges and are in a critical period of transformation and innovation. As platform providers, telecom operators need more intelligent and open business systems. One solution is to form new operation architecture by adding a core BDCS operation system to existing operation architecture. This allows resources to be provided in a distributed and virtual way. Operators should also integrate research and development, product manufacturing, network construction, application development, and service providing in order to establish more open ecosystem and business model.

References:

[1] Xiaoyu Tong, Yunyong Zhang and Yuanshun Dai, "Architecture and key technology of public computing communication network," *J. Communications*, vol. 31, no. 8. pp.134–140, 2010.

[2] Xiaoyu,Tong, G.Wu and Yunyong Zhang, *Post Telecommunications Age*, Beijing: Posts and Telecom Press, 2004. ch.2, p.27.

Biographies

Xiaoyu Tong (tongxy@chinaunicom.cn), is deputy director and senior engineer at the research institute of China Unicom. He has participated in several major ICT projects and has won several science and technology advancement prizes and management innovation prizes. He has taken the lead in R&D of the soft switching–based UniOne network and has developed the Sichuan Agricultural Information Network, which won the UN's World Summit Award. Key technologies and service mode in the Sichuan Agricultural Information Network have been widely applied in the telecommunications industry. He has published four books and 40 papers.

Yunyong Zhang (zhangyy@chinaunicom.cn), is deputy manager and senior engineer at the R&D department of China Unicom. His research interests include next–generation open networks, integrated fixed–mobile core networks, mobile Internet and services, and common computing. He has participated in projects of the General Armaments Department of the PLA, the Doctoral Fund of the Ministry of Education of China, National High–Tech R&D Program of China ("863" Program), National Basic Research Program of China ("973" Program), and the National Natural Science Foundation of China. He has published 15 books and 58 papers.

Bingyi Fang (fangby2@chinaunicom.cn), has a PhD in engineering. He is a senior engineer at the Research Institute of China Unicom and is mainly engaged in researching cloud computing and new core network technologies.

Roundup

ZTE Selected by BelTelecom for GPON National Broadband Network Project

ZTE Corporation announced on May 30, 2011 that it had won BelTelecom's national broadband network project in Belarus.

Under the contract, ZTE will use GPON technology to design and deploy a nationwide fiber–to–the–home (FTTH) network for BelTelecom. As a major component of BelTelecom's national broadband development strategy, this project aims to build a high–speed network that spans the entire country.

The project comprises multiple phases. The first covers all the seven regions of the country, including the capital of Minsk, with expansion planned for subsequent phases. Once completed, the project is expected to provide data and video services, as well as broadband applications such as distance education and telemedicine, across a high–speed Internet infrastructure. In addition, it will also deliver HDTV service to users across the country.

Belarus is a major member of the Commonwealth of Independent States (CIS) with a population of about 10 million. As the top fixed–line operator of Belarus, BelTelecom operates all the fixed–line and data transmission services in the country and holds a stake in the top 3 local mobile operators.

As of 2010, BelTelecom had five million users. ZTE started to exclusively provide BelTelecom with an entire suite of IPTV products in 2007, and has since supplied the operator with more than 1.5 million ADSL terminals. Building on the cooperative relationship, ZTE participated in the testing and bidding for BelTelecom's national broadband network project and was eventually selected as a key partner for FTTH construction.

ZTE's FTTH solution is based on the ZXA10 C300next–generation integrated access platform, the, which provides it with the largest switching capacity in the industry. It uses a diverse range of optical network terminals (ONTs) to meet user requirements under different scenarios. Its well–rounded EasyService O&M solution enables terminal management, fast service deployment and fault diagnosis across the network. With advantages such as environmental friendliness, smooth evolution, and easy O&M, ZTE's FTTH solution won BelTelecom's recognition and praise.

(ZTE Corporation)

The Internet of Things and Ubiquitous Intelligence (2)

Dongliang Xie
Yu Wang

(State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications)

Editor's Desk:

The traditional Internet is oriented towards person–to–person connection, whereas the Internet of Things (IoT) is oriented towards connections between inanimate objects. IoT covers a larger range of connections and involves more semantics. Traditional Internet and telecom networks focus on information transfer, but IoT focuses on information services. By combining sensor networks, Internet, telecom networks, and cloud computing platform, IoT can sense, recognize, affect, and control the physical world. The physical world can be unified with the virtual world and human perception. This lecture discusses IoT technology from three aspects: ubiquitous information sensing, ubiquitous network convergence, and intelligent information service. In this part, we discuss the architecture of sensor network and the status of the industry.

3 Sensor Network Architecture

Research on sensor network architecture is currently directed at the relationship between network nodes; that is, how the network is organized. In a traditional wireless sensor network (WSN), the nodes, which are large in number, often communicate with each other in a peer–to–peer, multihop, self–organizing manner to finish a user–specified task. With static, fixed data access nodes, the network has many inherent problems, including uneven energy consumption among nodes, low data transmission efficiency, inflexible deployment, and single network structure. It is also prone to route holes, coverage holes, and bottlenecks in nodes. These problems decrease the overall performance of the network. As a result, hierarchical WSNs have become a research area of interest in recent years. Compared with traditional WSNs, hierarchical WSNs optimize network performance in terms of energy efficiency, throughput, real time, reliability, and scalability [1].

3.1 Flat Architecture

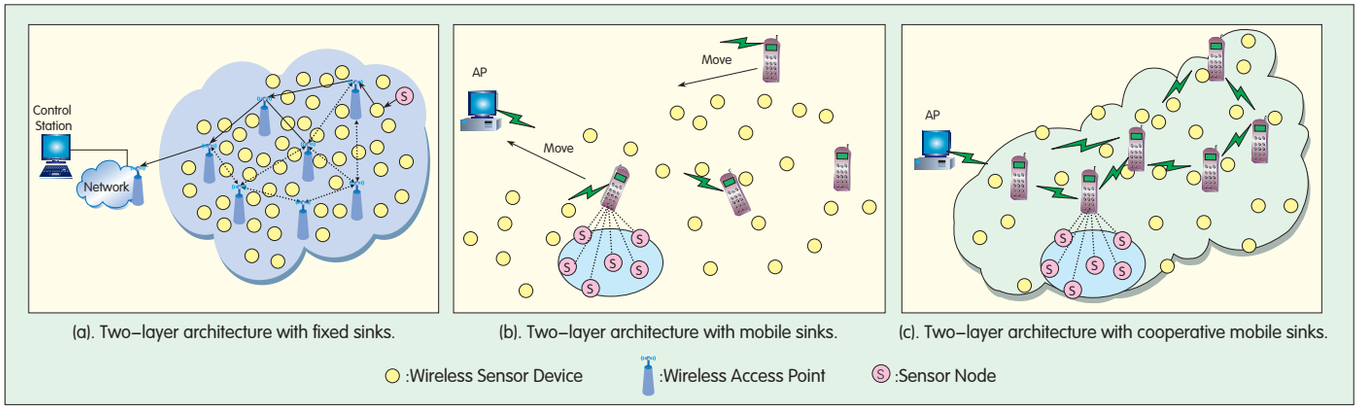
The flat WSN consists of a large number of static nodes that are distributed in a certain geographical area. In such a network, sensing data is transferred from the source node to remote sink nodes in a multihop and self–organizing way. Normally, these nodes have similar energy, storage, computing, and transmission capabilities, which means they are homogeneous. The data flow is multiple–to–one, and the neighboring nodes of the sink act as forwarding nodes in case of massive data. As the network size grows, some of these neighboring nodes inevitably become bottlenecks. Consequently, network performance decreases, and the network might even go down.

In a multihop network, growth in network size increases the possibility of data loss during transmission. It also increases the number of forwarding nodes, which, in turn, leads to sharp increase in energy consumption. Hence, network performance decreases as the network expands.

3.2 Two–Layer Architecture

So that nodes neighboring the sink do not become network bottlenecks in flat architecture, two–layer architecture is introduced into sensor networks [2]. Typically, a network with two–layer architecture selects a certain number of nodes as fixed access nodes (Fig. 1(a)). These fixed–access nodes, sparsely spread in the network, form an upper–layer coverage network to forward the node information of neighboring areas in a centralized manner. In this way, energy consumption within the nodes is balanced to some extent, and network performance is increased. However, near the fixed–access nodes, there are still some bottleneck areas with large energy consumption and heavy traffic.

Terminal technologies tend to be diverse, intelligent, and multimode. Portable electronic products such as mobile phones, notebooks, and PDAs, have powerful computing and communication capabilities and are highly mobile. They are beginning to replace fixed sinks in traditional WSNs and act as mobile sinks. So new sensor network architecture with mobile sinks is formed (Fig.1(b)). A mobile sink moves randomly within the network,



▲Figure 1. Two–layer architecture with sinks.

acquires the node information of neighboring areas, and forwards the data to access points. These mobile sinks can cooperate to form a self–organizing network (Fig.1(c)). Cooperation between mobile sinks can enhance the performance of WSN noticeably. With complicated processes such as data processing, access processing, data forwarding, and routing maintenance delegated to mobile terminals, the WSN minimizes data errors (or loss) arising from multihop transmission and uses the powerful computing capability of mobile terminals to share its information processing load.

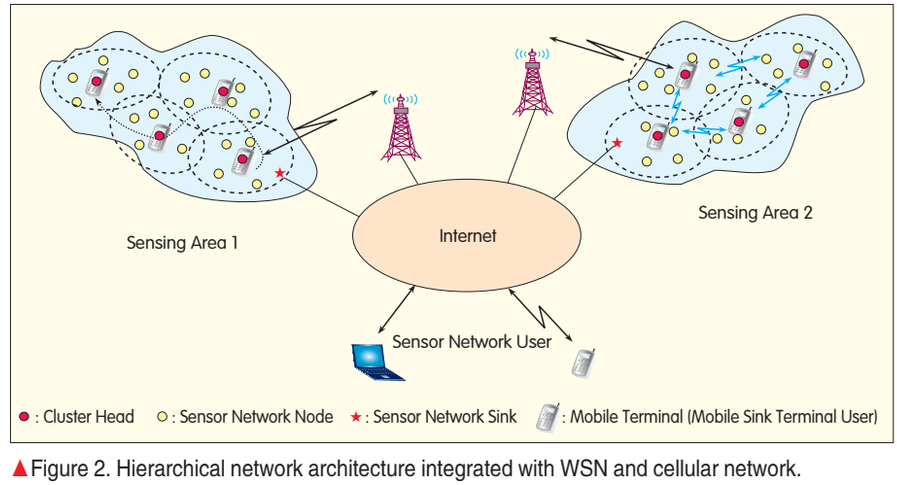
The two–layer architecture changes the data transmission mode of traditional flat architecture from multiple nodes to one fixed sink. It prolongs the system's lifetime, balances network energy consumption, increases the data transmission rate, and improves network coverage. But optimization of two–layer architecture is limited to the sensor network itself, and convergence of the WSN with other networks is not taken into account in the context of heterogeneous networks.

3.3 Three–Layer Architecture

Continuously evolving radio technologies provide a ubiquitous, heterogeneous network environment. Heterogeneous wireless networks have different background, objectives, development directions, system architecture, coverage ranges, communication protocols, link characteristics, application scenarios, and service provision capabilities [3]. As a result, three–layer sensor network architecture (Fig.2) is developed. Combining the infrastructure–based cellular network with the infrastructure–free sensor network, three–layer architecture makes full use of the complementary features of the two kinds of networks and solves network performance–1degradation caused by fixed access points. So it is quite suitable for future ubiquitous, heterogeneous, cooperative networks.

Specifically, the cellular network has a powerful service platform, mature operation mode, and sound management system, but its centralized control and management system makes it less flexible. Owing to its self–organizing feature, a sensor network is quite flexible, but its transmission distance is short and it lacks a mature operation mode and management system. Integration of the two networks enables a locally–deployed WSN to acquire information via the coverage of a mobile WAN, and to transfer and exchange the data in a wider range. Mobile WAN uses the rich information collected by the WSN to expand its service capabilities. Efficient exchange of information is achieved between machines, between machines and human beings, and between human beings and the real environment. The overall information handling process—from collection, transmission, and processing to reaction—is optimized. A harmonious connection is established between human beings and their surroundings.

Integration of WSN and mobile communication networks brings with it technical challenges. First, integrated networks and services comprising IP and non–IP technologies should make full use of multihop and self–organizing



▲Figure 2. Hierarchical network architecture integrated with WSN and cellular network.

features of infrastructure–free sensor networks in order to integrate with infrastructure–based networks. Second, with a research focus on the division, definition, and abstracting functions of layers and planes of ubiquitous heterogeneous networks, new communication network architecture models should be developed. These should be based on future network communication requirements as well as latest technologies in autonomous computing, autonomous communication, cognitive network, and ubiquitous computing. Then, diverse QoS demands of the Internet and new services in ubiquitous mobile networks can be satisfied. Third, diversity, intelligence, and multimode in terminal technologies should be exploited in order to study self–organizing, cooperative technologies for convergence of the sensor and mobile communication networks. Integration of terminal capabilities, communication methods, and access means at the stub area of networks can then be achieved.

The new hierarchical architecture integrates in a complementary manner the WSN (deployed in a specific area), the locally–deployed wireless self–organizing network, and the mobile network deployed in a wide area. Its heterogeneous link can improve data transmission rate and transmission reliability, and its heterogeneous energy can prolong the network's lifetime and improve the network's robustness.

3.4 Characteristics and Advantages of Hybrid Networks

The distinctive characteristics of a hybrid network are its heterogeneity and mobility. Heterogeneity in node energy, bandwidth, link, and computing capability improves energy efficiency in the node as well as throughput, reliability, and scalability in the network. It also expands the application scenarios of the WSN so that deployment of the WSN is easy. Mobility allows an agent device to dynamically mine information. This shortens the transmission link, reduces energy consumption, and alleviates an imbalance of energy distribution.

Mobility improves network performance dramatically [4]. A mobile agent maintains energy and prolongs the system's lifetime by reducing the traffic of each node. By decreasing the number of hops, the agent significantly decreases the probability of errors and increases the reliability of received data, which, in turn, reduces energy consumption caused by retransmission errors. When the mobile agent functions as a special relay node of the network, it can improve data transmission efficiency and reduce delay. Moreover, the mobile routing agent can solve the problem of non–connectivity in a sparse network. In a sparse, large–scale WSN, the relatively low density of sensor nodes often decreases connectivity in the network, and data transmission is affected. Using mobile nodes as mobile sinks for data collection can remedy the defects in a sparse network and improve the network's overall performance.

Heterogeneity in energy and link also creates many advantages. First, with enough high–energy nodes included in the network, multiple–to–one transmission bottlenecks can be solved. Second, data packets can reach sinks without being forwarded by low–energy nodes, so the network's lifetime is prolonged. Third, link heterogeneity reduces the average number of hops from the source node to the sink. The reliability of a sensor network link is relatively low, and each hop decreases the end–to–end transmission rate. A backbone link provides a cross–network high–speed link, which increases transmission rate and decreases energy consumption. Compared with common nodes, some mobile devices are more intelligent, programmable, and portable. With the popularization of mobile phones, mature, large–scale infrastructure has been developed in urban areas.

4 Research and Industry Status

In recent years, wireless communication technologies have developed rapidly. Wireless technologies provide users with a ubiquitous, heterogeneous network environment comprising wireless personal area network (WPAN) (such as Bluetooth), wireless local area network (WLAN) (such as Wi–Fi), wireless metropolitan area network (WMAN) (such as WiMAX), wireless wide area network (WWAN) (such as 2G and 3G networks), satellite network, Ad Hoc network, and WSN.

Each of these heterogeneous wireless networks has its own background, objectives, development direction, system architecture, coverage range, communication protocol, link characteristics, application scenarios, and service provision capabilities. These heterogeneous networks can be infrastructure–based or infrastructure–free. Infrastructure–based networks, represented by cellular mobile communication network and WLAN, have base stations, access points and routers. Infrastructure–free networks, represented by mobile Ad Hoc networks and WSN, are based on Ad Hoc technology and are dynamic, multihop, non–centralized, and self–organizing.

Infrastructure–based and infrastructure–free wireless networks are complementary. Infrastructure–based wireless networks have a powerful service platform, mature operation mode, and sound management system. But their centralized control and management system makes them less flexible. Because infrastructure–free networks are self–organizing, they are quite flexible. But their transmission distance is short, and they lack a mature operation mode and management system. Although these wireless networks provide users with diverse communication modes, various access means, and ubiquitous access services, they cannot deliver self–organizing, adaptive, ubiquitous services before they are truly integrated. Such integration involves cooperation based on complementary features.

At the same time, terminal technologies tend to be diverse, intelligent and multimode. Terminal types are becoming increasingly diverse. As well as traditional PCs and mobile phones, there are now sensor terminals with powerful sensing, computing, and communication capabilities as well as data terminals configured with Radio Frequency Identification (RFID) chips. With greater integration of control intelligence, these terminals have become intelligent information terminals not only capable of delivering multimedia voice, data and video services, but also capable of accessing the Internet for browsing, downloading, and online transactions. Overlapped coverage of various wireless networks is driving the rapid development of multimode terminals, and the development of terminal technologies makes it possible to support cooperative technologies for heterogeneous network convergence.

Research institutes, leading companies, and international standardization organizations see cooperative terminal technologies for heterogeneous network convergence as an important part of their research on next generation wireless communication networks. Much research has already been conducted and experiments performed.

In the area of cooperative terminal technologies for heterogeneous network convergence, many universities and research institutes have proposed their new network models and tried to address critical technical problems. These models include the Unified Cellular and Ad-Hoc Network (UCAN) architecture jointly proposed by Bell Labs and the University of California, Pervasive Ad-hoc Relaying for Cellular Systems (PARCelS) by Yale University, Mobile-Assisted Data Forwarding (MADF) by Stanford University, the Self-organizing Packet Radio Ad hoc Networks with Overlay (SOPRANO) project funded by the U.S. National Science Foundation, the hybrid network model, Sphinx, proposed by the Georgia Institute of Technology; Integrated Cellular and Ad Hoc Relaying systems (iCAR) by the State University of New York, Multi-Power Architecture for Packet Data Cellular Networks (MuPAC) and Throughput Enhanced Wireless in Local Loop (TWiLL) by the Indian Institute of Technology, and Multi-hop Cellular Network (MCN) by the National Chiao Tung University of Taiwan. These models combine the cellular network and mobile self-organizing network to form a hybrid wireless network. They aim for self-organizing relay and cooperation among terminals. The Using existing and future network infrastructures and the latest research achievements of Ad Hoc networks, the U2010 project aims to use cooperative technologies to provide the most capable means of communication and the most effective access to information during accidents, incidents, catastrophes, or crises. The Agent-Based Adaptive and Secure Service Provisioning for Mobile Users (ABASSMUS), Hybrid Wireless Network Communications (HyWercs), and Self-organized Network Infrastructures (SoNI) research projects of the University of Luxembourg are based on the hybrid architecture of backbone network and self-organizing network. They aim to enable mobile terminals to cooperate as service providers.

Leading companies have also studied the terminal cooperative technologies and have assisted operators in experiments with existing networks. They seek to optimize the performance of wireless communication networks and develop new service modes. The Symbiosis Institute of Business Management (SIBM) in India has proposed a Hybrid Wireless Network (HWN) model that supports multiple hops. Alcatel-Lucent has also launched the A-GSM network. These projects focus on the relay capability of next generation GSM networks, aiming to configure mobile stations with relay functions and change existing GSM systems as little as possible. This enhances the coverage of GSM networks.

International standardization organizations have done much research on cooperative technologies for heterogeneous networks. The ubiquitous network study group of the ITU proposes "ubiquitous network" and wireless technologies such as RFID with the aim of providing ubiquitous monitoring, sensing, and communication. Research has been done and trial networks have been established. The WWIF has proposed the Mobile Ubiquitous Service Environment (MUSE) model—a vision for future wireless communication. This model describes various cooperative technologies that may be used in future wireless networks at the service, network, and terminal level. In its TR25.924, 3GPP proposes Opportunity Driven Multiple Access (ODMA). 3GPP has studied relay cooperation, which supports mobile terminals, in the UMTS Terrestrial Radio Access-Time Division Duplex (UTRA-TDD) mode. On February 2, 2010, CCSA established TC10—the Ubiquitous Technical Committee— focused on ubiquity and omnipresence. (To be continued)

References

[1] R. C. Shah, S. Roy, S. Jain, and W. Brunette, "Data MULEs: Modeling a three-tier architecture for sparse sensor networks," *Proc. 1st IEEE Int. Workshop on Sensor Network Protocols and Applic.*, Anchorage, AK, 2003, pp. 30-41.
[2] L.Sankaranarayanan, G.Kramer and N.B. Mandayam, "Hierarchical sensor networks: capacity bounds and cooperative strategies using the multiple-access relay channel model," *1st Annu. IEEE Conf. on Sensor and Ad Hoc Commu. and Networks (SECON'04)*, Santa Clara, CA, 2004, pp.191-199.
[3] Biao Ren, Jian Ma, "mWSN: A Hybrid and Mobile Wireless Sensor Networks, " *1st Inter. Confer. on Mobile Compu., Commu. and Applic.* (ICMOCCA' 06), Seoul, Kroea, August,2006, pp.1085-1090.
[4] M. Yarvis et al, "Exploiting heterogeneity in sensor networks, " *IEEE INFOCOM 2005*, Miami, FL, March 2005, pp.878-890.

Biographies

Dongliang Xie (xiedl@bupt.edu.cn) is a director and associate professor at the Broadband Network Center of State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications. He researches wireless and mobile network technologies, wireless sensor networks, mobile Internet QoS, network cooperation, and ubiquitous intelligence. He has published more than 40 papers.

Yu Wang (wang0yu@gmail.com) is a master's student at the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications. She researches convergence of wireless sensor network and ubiquitous network.

Abbreviation Index

A

AAA: authentication, authorization, and accounting
ABASSMUS: Agent–Based Adaptive and Secure Service Provisioning for Mobile Users
ADC: analog–to–digital converter
AN: access network
AOEC: Auto Orthogonal Error Correction
API: application programming interface
AQEC: aerospace qualified electronic components
AS: autonomous systems

B

BDCS: business data center system
BGCF: breakout gateway control function
BI: business intelligence
BSS: business support systems

C

CA: core area
CD: cavity diameter
CH: correspondent host
CR: cognitive radio
CRM: customer relationship management
CSCF: call session control function

D

DAC: digital–to–analog converter
DHT: distributed hash table
DIOS: distributed intelligent open system
DNS: domain name server
DPD: digital predistortion
DR: dielectric resonator
DSA: dynamic spectrum allocation
DSM: delta–sigma modulator
DSP: digital signal processing

E

EID: endpoint identifier

F

FET: field–effect transistor
FIR: frequency–invariant reactance
FPGA: field–programmable gate array

G

GaN: gallium nitride

H

HIP: Host Identity Protocol
HIT: host identifier tag
HLR: home location register
HSS: home subscriber server
HWN: hybrid wireless network
HyWercs: hybrid wireless network communications

I

IBE: Identity–Based Encryption
iCAR: Integrated Cellular and Ad Hoc Relaying
IF: intermediate frequency
ISP: internet service provider
ITR: ingress tunnel router

L

LDMOS: laterally diffused metal oxide semiconductor
LISP: Locator/ID Separation Protocol
LNA: low noise amplifier

M

MADF: Mobile–Assisted Data Forwarding
MAPL: maximum allowed path loss
MBR: multiband radio
MCN: multi–hop cellular network
MEMS: micro electro mechanical systems
MGCF: media gateway control function
MIP: mobile IP
MuPAC: Multi–Power Architecture for Packet Data Cellular Networks
MUSE: Mobile Ubiquitous Service Environment

N

NAT: network address translation

O

ODMA: opportunity driven multiple access
ODS: open data services
OLAP: on–line analysis processing
OSS: operations support system

P

P2PSIP: peer–to–peer session initiation protocol
PA: power amplifier
PARCelS: Pervasive Ad–hoc Relaying for Cellular Systems
PCB: printed circuit board
PCCN: public computing communication network
PCN: public communication network
PID: packet identifier
PMIPv6: proxy MIP version 6
PON: P2PSIP overlay node
PRM: partner relationship management
PWPM: pulsewidth/position–modulation

R

RA: routing area
RAN: radio access networks
RTP: real–time transport protocol
RVS: RendezVous server

S

SAW: surface acoustic wave
SBC: session border control
SDC: subscriber data center
SDP: service delivery platform
SDR: software defined radio
SIBM: Symbiosis Institute of Business Management
SMPA: switched mode PA
SMW: Surface Mounted Multiband
SNR: signal–to–noise ratio
SoNI: Self–organized Network Infrastructures
SOPRANO: Self–Organizing Packet Radio Ad hoc Networks with Overlay
SQL: structured query language
SSO: single sign–on
STUN: session traversal utilities for NAT

T

TPA: third party auditor
TURN: traversal using relay NAT
TWiLL: throughput enhanced wireless in local loop
TZ: transmission zero

U

UCAN: Unified Cellular and Ad–Hoc Network
UE: User Equipment
UMTS: Universal Mobile Telephone System
UTRA–TDD: UMTS Terrestrial Radio Access–Time Division Duplex

V

VCC: voice call continuity
VSWR: voltage standing wave ratio

W

WBG: wide bandgap
WLAN: wireless local area network
WPAN: wireless personal area network
WSN: wireless sensor network
WWAN: wireless wide area network

Z

ZIF: zero intermediate frequency