REVIEW PAPER

Trends of Communication Processors

LIU Dake, CAI Zhaoyun*, WANG Wei

Lab of Application Specific Instruction-Set Processors, Beijing Institute of Technology, Beijing 100081, China

Abstract: Processors have been playing important roles in both communication infrastructure systems and terminals. In this paper, both application specific and general purpose processors for communications are discussed including the roles, the history, the current situations, and the trends. One trend is that ASIPs (Application Specific Instruction-set Processors) are taking over ASICs (Application Specific Integrated Circuits) because of the increasing needs both on performance and compatibility of multi-modes. The trend opened opportunities for researchers crossing the boundary between communications and computer architecture. Another trend is the serverlization, i.e., more infrastructure equipments are replaced by servers. The trend opened opportunities for researchers working towards high performance computing for communication, such as research on communication algorithm kernels and real time programming methods on servers.

Keywords: ASIP; baseband processor; network processor; application processor; server processor

I. INTRODUCTION

1.1 The motivation of the paper

There are different definitions on communication systems. From hardware point of view, equipment in a communication system is dominated by processors running protocols and surrounded with peripheral circuits. To get sufficient performance with adequate flexibility in an application domain, under constrained power consumption and silicon cost, Application Specific Instruction-set Processors (ASIPs) have been designed for critical parts of communication systems. Typical examples are radio baseband processors for mobile phones and network processors for core routers. Following the trends of SDN (Software Defined Network), classical infrastructure equipments are gradually replaced by general purpose computing servers. Processors have been serving an important role in communications, and it is therefore necessary to review processors designed for and used in communication systems.

The 1st dedicated communication processor might be the one designed for the first electronic switch machine in 1964 by Bell Labs [1], though it was not given a specific name. The first DSP (Digital Signal Processor) might be introduced to handle PCM switching in the first PCM machine in 1971 [2]. Following market needs, processor design methods have been paid attention to. The first processor generator MIMOLA was reported by Zimmermann and Marwedel in 1979 [3]. After that, there have been many ASIP generators (Cathedral, Target, LISA, and NoGAP) supporting designs of communications processors.

It has been proven by tracking the history of telecommunication, that the developments of modern communication systems are aligned with the development of communication processors. In each state-of-the-art communication system today, there are several processors and the communication system trends will be based on ASIP, server central processor, and AP (application processor). It is thus of essential to have a review and discussion on processors for communications by this paper.

1.2 The scope of the paper

A communication system transfer data from one place to another. It usually does not produce and consume data (payload). Most payloads are digitalized; their transmissions and controls are based on digital system maneuvered by processors. In this paper, we will focus on processors for transferring and controlling of data. A communication system can be partitioned into subsystems in the following Figure 1. The partition and classification of communication processors is also matched in Figure 1. Communication processors are allocated in the physical layer (PHY), such as radio modems and access modems; in the media access layer (MAC) such as the radio MAC controller; in the layer 2 and 3(L2/L3) such as radio protocol processing machines or IP protocol processing machines. In communication infrastructure, application processor is only designed for data format exchanges in gateway. In a terminal, an application processor is designed for both data format exchange and for generating/consuming data.

1.3 Wireless network

Wireless network consists of its infrastructure and terminals from 2G (GSM, CDMA)， 3G (WCDMA, CDMA2000, and TD-SCDMA) to 4G (TD/FD-LTE) for public services. There are also other wireless systems, such as trunked radio for services to specific domains. On the infrastructure side, there are base stations (including relays) as the majority part of the infrastructure.

In a base station, the baseband processor (ASIP or/and FPGA) takes the heaviest computing load for channel coding/decoding, channel equalization, filtering, and error corrections. High-end microcontroller units (MCUs) take computing load for MAC and L2/L3 protocols. Network processors (NP, as ASIP or ASIC) are used for backhaul accesses.

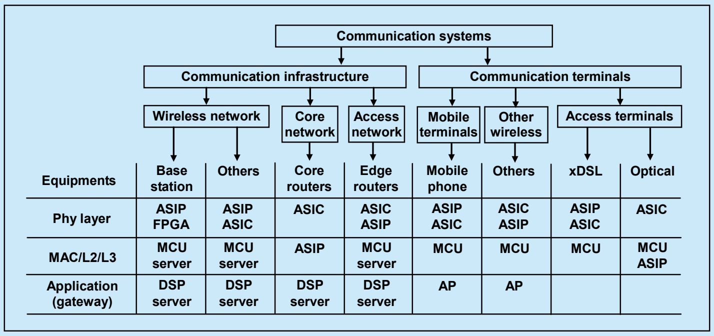
In infrastructure, there are also mobile switching centers for switching and services; and there are gateways connecting to other networks. MCU and NP are in mobile switching centers. In a gateway, there are DSP in the data path for payload format transcoding， and MCU/NP in the control path for inter-protocol message passing and processing.

On the terminal side, there are mobile phone and other terminal transceivers. Processors in a terminal include radio baseband processors (mostly ASIP), protocol processors (MCU) and application processors. Heavy communication payloads are generated from and consumed by AP in a phone. Discussion on AP might be out of the scope of the paper. However, as AP is a processor inside a communication device, we will offer a brief discussion and review.

1.4 Access network

By definition, access network offers data transfer for the last mile. It consists of Internet Service Provider (ISP) networks and supply network to network users outside an ISP. Inside an ISP, the internet is offered by Ethernet network for an enterprise or a group of people. Supply network services outside an ISP are

> Nowadays, processors have been playing important roles in both communication infrastructure systems and terminals. In this paper, both application specific and general purpose processors for communications are discussed including the roles, the history, the current situations, and the trends.



Equipments	Base station	Others	Core routers	Edge routers	Mobile phone	Others	xDSL	Optical
Phy layer	ASIP FPGA	ASIP ASIC	ASIC	ASIC ASIP	ASIP ASIC	ASIC ASIP	ASIP ASIC	ASIC
MAC/L2/L3	MCU server	MCU server	ASIP	MCU server	MCU	MCU	MCU	MCU ASIP
Application (gateway)	DSP server	DSP server	DSP server	DSP server	AP	AP		

Fig.1 *Partition of communication systems*

usually offered by xDSL (Digital subscriber line) and xPON (Passive Optical Network). Old access networks such as ISDN (Integrated Services for Digital Network) were expired and will not be discussed in this paper.

Edge routers are placed at the edge of an ISP network. Inside an edge router, Ethernet protocol processing is usually allocated in PHY ASIC chips. Edge routers are mostly designed based on specific network processors. TCP offload engine is designed using NP if the bandwidth is high, otherwise, a server could be used as a low bandwidth TCP offload engine.

Because server performance keeps going higher and cost of service keeps going lower, more edge routers are replaced by servers. When the edge bandwidth is high, gateway for VoIP (voice over IP) is usually designed based on ASIP-DSP. When the edge bandwidth is not so high, a server could be used as a gateway. When the edge bandwidth is high, a dedicated intrusion detection machine is used together with a firewall. Special ASIP or FPGA is used as the intrusion detection engine.

On the infrastructure side in xDSL machine at local access center, ASIP and FPGA are used for the PHY modem, and NPs are used as protocol processing engines. On the terminal side, modems are designed based on ASIP or ASIC and a MCU is used as the protocol processer. The PHY part, the SerDes (Serializer/Deserializer), could be in NP or a separate chip.

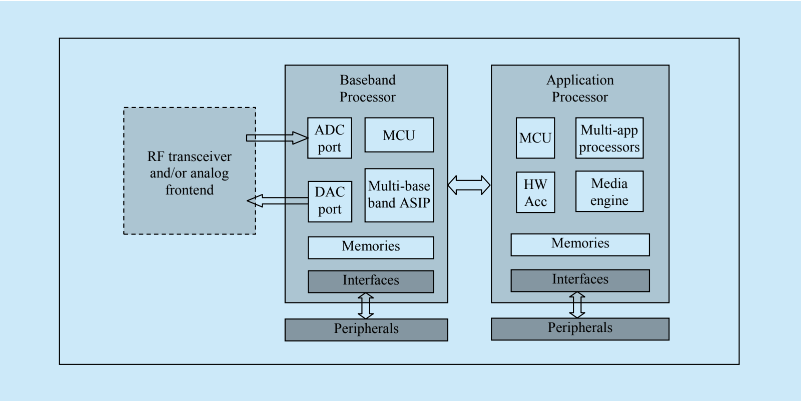


Fig.2 *Core IC in a communication device*

1.5 Core network

Both wireless and access networks are connected to core networks for long distance broadband data transmission. Core router plays the kernel role in a core network. When the port rate in a core router is not at the high-end (not more than 10Gbps in a port), NP are used for core routing. Otherwise, when the port rate in a core router is at the high-end (more than 10Gbps in a port), both ASIC and NP can be used for core routing.

II. Baseband Processors (BP: Modems)

The core integrated circuit in a communication device is depicted in Figure 2. It is consisted of 3 main parts: the analog front-end (including RF transceiver), the baseband processor and the application processor [4].

Baseband processor (BP) is the front-end processor managing all transmission and receiving functions. Application processor is the back-end processor aiming at supporting a wide range of user applications for communication devices, including voice, image, audio, video, and 2D/3D graphics.

Mobile communication processors can be divided into 3 categories: Stand-alone BP and AP; communication processors with integrated BP and AP; and integrated baseband and RF transceiver [5]. A physical layer BP can be in a mobile terminal, in a base station, and in an access network facility.

2.1 BP Function Review

Typical OFDMA baseband transceiver (transmitter and receiver) function flow is in Figure 3 [6]. There are 4 functional modules: symbol processing, bit processing, FEC, and MCU. In a receiver, functions are:

Digital front-end (DFE): Matching sampling rate and carrier frequency offset (CFO) using decimator, rotator, and farrow filter; band-pass filters for band selection and noise control; and gain control.

Synchronization: Synchronizing bits and frames based on cyclic prefix and preambles using cross correlation or auto correlation.

Transform: Transforming signals between time and frequency domains for OFDM modulation and for cross-domain processing.

Channel estimation and equalization: Channel estimation is to get channel state information (CSI). CSI is then used to compensate for signal distortions in the channel equalization stage.

Detection: Equalized symbols are decoded to soft/hard bits through detection algorithms, such as log likelihood ratio (LLR) demapping.

De-interleaving: Interleaving shuffles data bits with sufficient distance, to avoid bit damage by glitches. De-interleaving is the inverse process of interleaving.

Forward error correction (FEC): Redundant codes are added on transmitter part and are used at receiver part for correcting errors induced during transmission.

Cyclic redundancy check (CRC): A remainder of a polynomial division generated at the transmitter part can be used as redundant part for error detection. CRC is thus on both parts.

In a transmitter, not yet mentioned functions are:

Rate matching: To match the data block size to the radio frame size by filling or puncturing off bits.

Precoding: for pre-distortion, training, or channel adaptation. It carries special features on payloads.

Modulation: Mapping digital bit-streams onto carrier waveforms for physically transmission. OFDM modulation merges multiple modulated subcarriers into a long one.

Pulse shaping filtering: Changing the transmitted waveform to minimize inter-symbol interference (ISI) and interferences to other channels.

Digital pre-distortion (DPD): To correct distortions and improve the linearity of antenna power amplifier in radio frequency (RFPA).

2.2 BPs in different equipments

2.2.1 BP in wireless terminal

BP in a wireless terminal needs the computing performance up to several hundreds of giga operations per second, while the power budget is limited to only 1W. In addition, a modern BP in a smartphone shall support multiple mobile communication standards, such as GSM/EDGE, WCDMA, HSPA, LTE, LTE-A, and WLAN with low silicon cost. The trade-off is thus essential among performance, power and area consumption, and degree of flexibility.

To meet the above requirements, low power DSPs or ASIPs are adopted as baseband multiprocessors. An architectural example is the heterogeneous ASIP architecture [6] based on software-defined radio (SDR). A top-level view of the architecture is in Figure 4. The baseband MCU offers controlling for all ASIP, ASIC, and interfaces in order to correctly conduct data processing flow in the baseband subsystem. ARM architecture is usually used. Each of the heterogeneous application-specific coprocessors is optimized to reach computing and memory (access) efficiencies in a specific domain. They usually employ VLIW and SIMD architecture to accelerate baseband kernels with high degree of instruction-level and data- level parallelism. ASIPs are used to provide high throughput with ASIC-like power consumption.

2.2.2 BP in base station

In a base station, a BP takes baseband process-

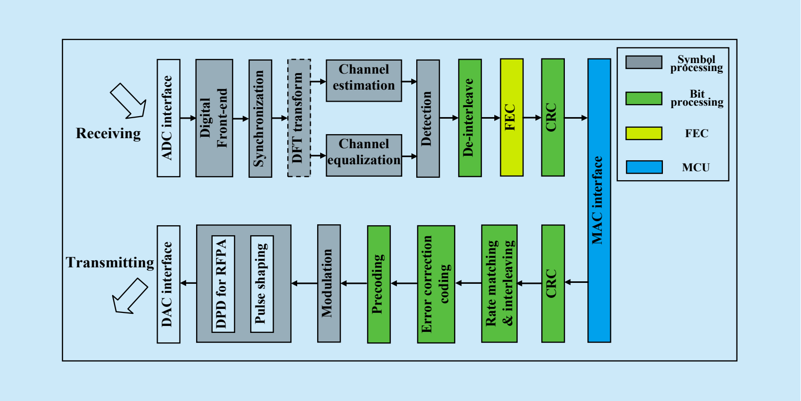


Fig.3 *Baseband function flow*

ing for multi-user. Tasks are either user-shared or user-specific [6]. Cross-interference cancellation shall be taken into account between users.

Advanced processing techniques, such as adaptive modulation, space-time coding, beamforming and MIMO antennas are introduced to enhance the system performance. Power consumption is not as critical as that in terminals, whereas the infrastructure cost is of great concern for manufacturers and operators. BP in a base station has to be flexible and scalable for various wireless standards and configurations.

BP in base station usually employs a SDR platform with a mix of FPGAs and DSP/ASIPs. DSP/ASIPs provide high speed signal processing, and are mainly used for encryption/decryption and modulation/ demodulation. FPGAs can provide better flexibility for programming-hard algorithms as the complement of DSP/ASIPs. The algorithm selection, partition and structure design must offer configuration freedom for number of users and flexible bandwidth.

2.2.3 BP in access network

There are access networks, such as transmission through twisted pair, coaxial cable, power cable, and optical fiber. BPs in access network are similar to those in radio systems, the main difference is the longer CSI life time. The period of channel estimation is thus longer. In this paper, we will merge the baseband processors in access network infrastructures and terminals, because the main difference is the count of channels. Even though upstream receivers in an infrastructure may have co-process for cross interference of MIMO channels.

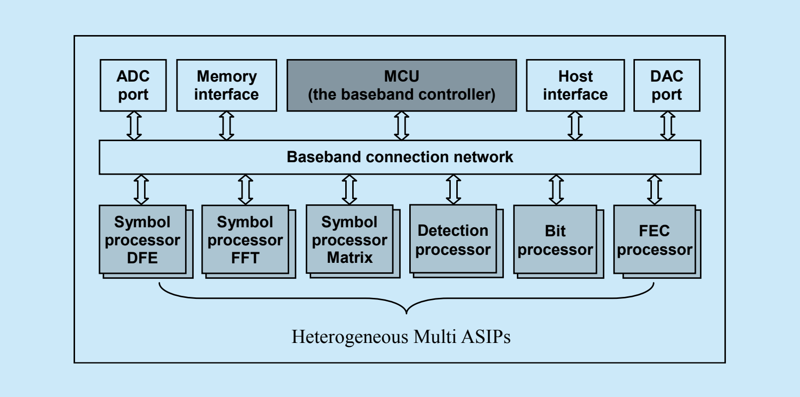


Fig.4 *Heterogeneous ASIP architecture for SDR BP*

BP in access network may vary according to the transmission technologies. Popular technologies are xDSL, power line communication (PLC), and PON.

xDSL is a technology family transmitting digital signal through telephone lines with different modulation schemes. For example, Asymmetric Digital Subscriber Line (ADSL) [7] uses Discrete multi-tone (DMT) modulation scheme. BP for ADSL handles DMT transceiver functions including error checking/correction, convolutional coding, decoding, mapping/demapping, FFT/IFFT, etc. Current baseband processing is implemented by DSPs and ASIPs for the various xDSL standards.

PLC uses power cable wires to simultaneously conduct data transmission and electric power of Alternating Current (AC). PLC is used for different applications, such as home automation and Internet access. When used for high-speed Internet access, it is also known as Broadband over Power Lines (BPL), which is standardized by IEEE 1901 [8]. OFDM is used for uplink and downlink transmission in this standard. In narrow band PLC for control/home automation applications, BP for Direct Sequence Spread Spectrum (DSSS) scheme is needed [9].

PON is a telecommunication network using optical fiber and unpowered optical splitters to achieve point- to-multipoint transmission. The ITU-T G.984 Gigabit -capable Passive Optical Networks (GPON) [10] and IEEE 802.3ah Ethernet PON (EPON) [11] are the most widely applied PON standards for broadband access, and 10G-PON is the next generation ultra-fast network technology. There is no BP and an ASIP for SerDes is needed.

2.3 Discussion on BP

2.3.1 History

Early wireless and access devices were designed to support only a single standard. Their BPs were mostly ASIC-centered and DSP-assisted [12]. ASIC chips offer high performance over power consumption, and are suitable for the processing of a single standard. However, as multiple communication standards appear, an ASIC has to integrate multi hardware modules to support multiple functions and algorithms, which causes significant increase of hardware cost and complexity. In 2001-2005, baseband solutions based on SDR were proposed by NXP, Icera, Coresonic, and Sandbridge. SDR became popular. Several alternative approaches have then emerged: reconfigurable architectures, such as ADRES [13] by IMEC; ASIP/DSP-centered architectures, such as Leocore [14] by Coresonic and Sandblaster [15] by Sandbridge; and multi-core architectures, such as SODA [16] and Tomahawk MPSoC [17]. All of these architectures offer a certain degree of flexibility as well as efficiency for multiple mode processing. At the same time, baseband ASIPs [18] have taken the place of traditional DSPs. Baseband ASIPs are optimized in instruction set micro-architecture for only baseband functions, e.g., filtering, FFT, matrix, and bit processing. Parallel computing architectures such as SIMD and VLIW are also widely adopted by baseband ASIPs.

In recent years, an ever-increasing demand for data transmission services have been addressed by wireless and access standards, such as 4G/LTE and LTE-A. For wireless network, this demand is also addressed by deploying small cells in areas of high traffic volume, such as residential areas, enterprises, shopping malls, etc., thanks to the introduction of Heterogeneous Network (HetNet) technology [19]. It incorporates base stations of various kinds, such as femto cells, metro and pico cells, and uses wireless backhaul to offload data from a macro base station. A small cell BS is required to offer an increased network capacity. Its BP has to offer high computing power, yet must be very cost-effective. Many companies have expanded their business to support small cell deployments, including new BP chip-set and SoC solutions. Available products include Broadcom's BCM617xx [20], TI's Keystone II [21], Qualcomm's FSM99xx [22], etc.

2.3.2 Trends of BP

Silicon technology scaling will no longer provide the power benefit as it once did [23]. The solutions for now and the future are in 2 aspects: multi-core and ASIP. Multiple small cores at a lower clock frequency are more power efficient, and can help to exploit parallelism at various granularity levels [24]. ASIP can offer high throughput and computing capability while maintaining power and cost efficiency.

System-on-chip using multi-processor (MPSoC) is a promising solution for mobile devices that replaces the traditional baseband architecture style. Multiple cores are implemented on a SoC for control and data processing with increased power efficiency. There are homogeneous MPSoCs that employ replicate processor cores or configurable hardware, such as the coarse grained array (CGA) architecture of ADRES, which implements 16 interconnected functional units. There are also heterogeneous ones which employ specially optimized processor cores and dedicated hardware accelerators for certain class of functions. Tomahawk2 [25] is a recent approach. The key aspect of efficiency is to build a programming model that fully utilize the hardware feature of multiple cores and its on-chip interconnection network, thus the overhead of computing, task scheduling and communication can be minimized.

ASIP architecture is efficient for BP in its flexibility and performance over power or silicon cost. In state-of-the-art BP approaches [26], VLIW architectures combined with SIMD architectures of wide datapath (up to 32 lanes and 512-bit data width) have been adopted. ASIP instruction set tend to fuse more basic arithmetic operations in one instruction, or utilize hardware accelerators for specific algorithms, thus to minimize the overhead in innermost loops and data moving [27]. However, increased parallelism and novel instruction set can be at the expense of complex addressing, sophisticated memory access, as well as tough programming. The design of addressing path, memory subsystem, programming and compiling will be of significant consideration for researchers.

III. Network Processors

The network speed keeps increasing and challenging to infrastructures (switches, routers, gateways, and firewalls). These network equipments offer complex and intensive processing on each data packet at real time or line speed. Network processors are the engines of network equipments and are usually ASIP. Here the ASIP stands for NP designed delicately for packet processing with excessive higher performance comparing to general purpose CPUs. Programmability makes NP an easy and flexible solution to meet diverse protocols and scenarios.

3.1 NP function review

There are three key aspects of packet processing functions in a NP: path, direction, and specific tasks [28]. Figure 5 illustrates the general framework of packet processing flow. Packet processing follows two paths:

Control path (slow path): for upper level processing, e.g., routing table update and routing decisions.

Data path (fast path): for the network processor functions in a specified sequence.

Generally, the processing flow in Figure 5 can be separated to 2 directions, which are:

Ingress: functions related to entering the equipment or the network processor, from the network.

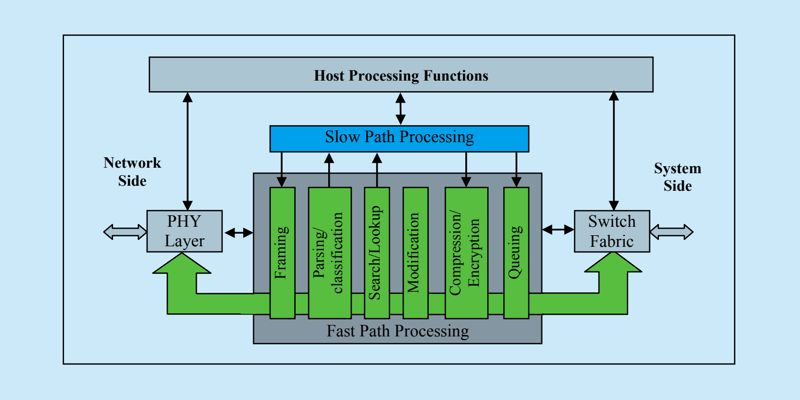


Fig.5 *A general framework of packet processing*

Egress: functions related to exiting the equipment or the network processor, to the network.

In a NP, the ingress and egress functions might not be explicitly separated in architecture. However, the function flow is actually different. In Figure 6(a), ingress and egress are separately implemented in two half-duplex NPs. In the upper part (1) there are line cards for receiving packets from and transmitting packets to the network, while part (2) consists of switch fabric, service cards, and other forwarding and processing mechanisms. In the ingress direction, the packets enter from the left, and go through the fast path controlled by slow path (as illustrated in Figure 5); the packets are then forwarded either to a switch fabric or to the network (line interface) again to the egress direction. Figure 6(b) shows a different scheme, which is the full-duplex architecture. Both ingress and egress functions can be performed in one processing core.

Packet processing tasks include:

Framing: In fast path, framing tasks include: synchronization, CRC error detection and integrity check in the ingress process; segmentation /fragmentation, assembly and error detection and correction coding in the egress process.

Parsing and classification：In fast path of ingress, parsing is the first analysis to synchronize, identify and extract the fields in an incoming packet, error check, and to decide for further packet processes.

Packet classification is one of the main tasks in NP. It classifies packets into flows to be processed by NPs. These flows are defined by rules of protocols. The rule database contains many entries, each of which is composed of a pair of a specific rule and its action. Each incoming packet only matches with specific rules, and the appropriate action can be taken on the packet. For example, in packet forwarding, the rule is the packet destination address, and the action is to forward the packet to its destination. Classification can be implemented in several ways. Software classification enables dynamic classification criteria. Simple software cannot support classification at high wire speed. However, hardware classification can be fast yet inflexible.

Search, (IP-)lookup, and forwarding: Search, or lookup, is an essential and mostly called task in fast path. One example is to match an IP address in a table. Searches through huge tables at wire speed are time-critical and power-hungry. Search engines can be in software or hardware (a programmable functional unit). Typical hardware approach is based on Content Addressable Memory (CAM), who can return the address of the matched content value in a clock cycle. CAMs are divided into regular CAMs (or binary CAMs) and ternary ones. By adding “don’t care” in search, TCAMs allow masked searches, which is of great advantage in IP lookup applications.

IP lookup is the most frequently executed function in packet forwarding, and is also required by classifications, billing, access lists, and other applications. Packet forwarding has to rely on rapid IP lookup to make fast decisions on a destination egress port. Thus, in high-speed networks, acceleration of IP lookups is essential to any network processor design.

Modification: Packet modification includes content changing, deleting/adding, canceling the entire packet, and duplicating a packet.

Compression and encryption: Compression and encryption might be executed at a network edge, or in access networks. Coprocessors or hardware modules are usually attached to NP to perform the compression and encryption task.

Queuing and traffic management: Queuing and traffic management are on packets in sending queues. It schedules sending packets according to line and resources, and decides packet parameters, such as priority. It is usually implemented in slow path.

3.2 Different network processors

3.2.1 Core router NP

A core router for trunk or core network supports routing/switching of multi ports at wire-speed up to IP layer with port rate of 10–100 Gbps.

Hardware implementation of core routers can be a mixture of ASIC and ASIP. ASICs are used for Ethernet PHY and MAC components; ASIPs are responsible for upper-layer packet processing. ASIPs are optimized in architecture and instruction set to efficiently run packet processing tasks. Heterogeneous NP architecture is usually based on multi-core parallel and pipelined 2D array with a dedicated task-specific instruction set.

3.2.2 Edge router NP

Edge routers are placed at the edge of an ISP network connecting an ISP to core network. An edge router handles protocol processing from IP level up to application levels. Edge router NP is mostly designed based on specific network processors.

Currently, as server performance keep going higher and cost going lower, more edge routers are going to be replaced by servers, such as Broadband Network Gateway (BNG) which including Broadband Remote Access Server (BRAS) functionalities. The trend of “serverlization” will be discussed in section 5.

3.2.3 Protocol MCU

Protocol MCU can be identified from equip-

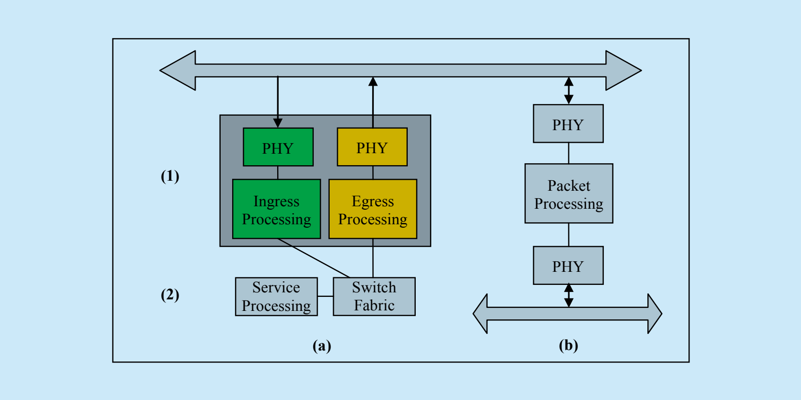


Fig.6 *Network processor implementations*

ments such as PSTN switching machine (for Signaling System No. 7—SS7), mobile switching center, and access devices such as xDSL, PLC, xPON, as well as edge router.

In an infrastructure machine, high-end MCU such as PowerPC, ARM and MIPS are commonly used for protocol processing. For radio and access terminals, low-end products of ARM and MIPS are employed.

3.3 Discussion on NP

3.3.1 History

In 1990s, most network processing was ASIC-based. The reason is that general purpose CPU is far behind the wire-speed requirements, while ASICs hold the best performance and lowest power consumption. NPs appeared in the late of 1990s, and has been based on parallel architecture due to the growing gap between the network speed and processor speeds [28][29](i.e., networking requirements always increase faster than Moore's law [30][31]). Symmetric Multiprocessor architecture is extensively adopted in first-generation NPs, such as IBM PowerNP[32] and Intel IXP[33]. After 1999, pipelined homogenous processor architecture (Xelerated NPs[34]), pipelined heterogeneous processor architecture (Agere NPs[35]), parallel pipelines of processors (Cisco Toaster [36], Freescale C-Port NPs [37]) and pipeline of parallel processors (EZchip NPs[38]) were employed. At the same time, much protocol processing tasks are also done in CPU supported by SMP (symmetric multiprocessing), e.g., IBM PowerPC (the first architecture of IBM enter into SMP world) [28].

Today, the line-speed in core network can range from 100Gbps to 400Gbps [39]. Under the Ethernet physical layer evolutions, the line-speed could be up to 1Tbps. The complex packet processing has to be completed in extremely short time period, e.g., tens of nanoseconds for framing, parsing, classification, search, lookup, modification and forwarding. Network processing at wire-speed is thus the greatest challenge for core network equipment. State-of-the-art network processors of this kind include Broadcom Sahasra™ Processor [40] and Cisco Flow Processors [41].

3.3.2 Trends of NP

For the future core networks, Ethernet PHY and MAC components will still be implemented with ASIC as the flexibility requirement is low. On the other hand, upper-layer packet processing shall be in NP ASIP. Support for multiple protocols will be the essential requirement in core network. NPs have to offer programmability for various protocols and for adapting new protocols. Processing for multi-protocol is up to the 7th layer. However, payload and upper layer processing (after TCP offload engine) are in edge router and is out of the scope of discussion in this paper, whereas MPLS for core router up to IP has to be necessarily considered.

Real-time networks, such as real-time bus in robot and industrial control systems need NP for framing, parsing and classification, searching and extraction, modification, and forwarding in line-speed with low power and low cost.

For example, EtherCAT [42] (Ethernet for Control Automation Technology) is a real-time industrial Ethernet standard for industrial applications. It is based on Ethernet with added synchronous real-time features and is fully conforming to the Ethernet specifications. To offer real time network, a small NP in EtherCAT shall get (extract) and send (insert) real-time data from/to each motion control unit and motor control unit with jitter less than few microseconds. Normal CPU cannot offer such synchronous functions for multi-ports. Similar requirements and NPs are also needed by PROFINET, EtherNet/IP, Powerlink, and SERCOS.

IV. Application Processors

4.1 Introduction to AP

AP, a modern backend processor defined in Figure 2, is usually a system-on-chip (SoC) design, to provide an operating system environment (e.g., android/ios, FreeBSD Linux) that supports diverse applications using payloads from frontend or NP. It is an absolutely indispensable part in a terminal (such as smartphone, automotive driving assistant, tablet device, etc.) and an infrastructure (primarily gateway) equipment to provide high quality of service to consumers, and to meet wide variety of theirs needs.

4.2 AP in terminals and infrastructure

4.2.1 AP in terminals

In user terminals, an AP may consist of a MCU (e.g. ARM Cortex-A72 based on ARMv8-A), a vector processor (e.g. Mali GPU), and a DSP (e.g. Mali Video or CEVA X). Both real-time and non-real-time tasks run on AP. Real time tasks are voice, audio, image, video, and games. Non-real-time tasks include web browsing, internet APPs, and cryptograph etc. Multicore AP is needed in a high-end Smartphone. The system architecture is usually mixed with SMP and HSA (heterogeneous system architecture). Inside, each core, accelerations can be found for video (video pre-processing (ISP), video CODEC, and video post processing), Encryption, data compression, Java, and vector computing. Cores are connected as a SMP processor or a HSA multi-processor via cache coherent interconnections [43].

4.2.2 Gateway processors in infrastructure

Gateway, or protocol converter, is a translation machine interfacing between different communication systems, such as a mobile gateway translates mobile control protocols and voice formats of radio access network to other networks', e.g., PSTN (public switched telephone network) or VoIP system.

In a gateway, there are normal MCUs for translating signaling and protocols (such as SIP (3GPP) and H.323 (ITU)). MCU is a general purpose processor and will not be discussed in detail in this paper. Voices in different systems are compressed based on different standards. In mobile systems, voice is compressed based on AMR (Adaptive Multi-Rate) and it shall be translated to ITU VoIP format (G.723, G.726, and G.729) when a mobile phone and an IP phone are calling each other. When a PSTN phone and other phone are connected, the line echo from PSTN local access line shall be canceled following ITU G.168, the voice compression format shall be between G.711 and VoIP or AMR.

A gateway data path shall thus handle both voice compression format translation and echo cancellation for channels as many as possible at the same time using high-end DSP processors. One voice translation needs about 5MOPS (million operations per second) for decoding and 25MOPS for encoding. One echo canceller for an 8kHz sampling channel needs about 30MOPS. Thus, the peak performance needed for a 512 connecting mobile-VoIP lines are more than 16GOPS. ASIP DSP for gateway is thus needed to accelerate the processing for voice CODEC and echo cancellers.

Gateway can also be allocated in an access terminal. In this case, a normal high-end MCU can be used for both the control path and data path.

4.3 Discussion on AP

4.3.1 History

Historically AP is mainly used for voice CODEC and MP3, such as ARM-based Cirrus Logic EP7209 [44], due to simple operating systems and limited applications run on AP in terminal side, e.g., feature phones. Later, the design and deployment of AP has experienced tremendous development affected by rapidly growing mobile communication industry in the last 20 years. Compared with 2G, 3G, data rates in 4G are much higher reaching up to 100+ Mbps and it will be even more in 5G, leading to higher demands for delivering experience like HD 1080p video, 3D stereoscopic image. Now demands are shifting to computer vision, virtual reality and other Big Data applications. These real-time and intensive tasks pose big challenges for APs to provide high computing capabilities without compromising to performance, power consumption and thermal budget in mobile devices. ASIP, such as video engine (NEON), Java engine (Jazella), compression engine, and encryption engines (TrustZone), are needed to reach acceptable performance and adequate programmability.

Challenges also boost technology progress, e.g., Accelerated Processing Unit (APU). APU represents a kind of fusion that integrates ARM cores, accelerators (or co-processors), and other analog circuit functions, such as memory controllers, USB host controllers, and DDR controllers as well as custom accelerators.

4.3.2 Trends of AP

In terminals, there are new demands like 4K video, 120fps camera, 3D online games, real-time object recognition, immersive multimedia and other deep learning applications. It needs more computing power for application processors without sacrificing the battery life and thermal dissipation envelope. It is thus predictable that heterogeneous computing capacity will be further enhanced. In hardware side, because the silicon scaling cannot follow the Moore's law, more efficient processors/cores, ASIP engines, and more efficient SoC interconnection architectures are needed. In software side, programming on HSA is tougher and needs advanced high level programming tools; kernels, for different algorithms running on dedicated ASIPs, are needed for intrinsic based programming.

Application processors in gateway or in other infrastructure devices, mainly, data-plane (voice CODEC, echo canceller, and video transcoding), would experience a similar development process to terminals. As mentioned trends in network processor, continuous growth in demand for high level computational capacity is accompanied also by higher needs for quality of service (especially latency) that pushes APs in infrastructure devices to adapt to novel technology, like HSA in server domain and multicore architecture. Server performance are going to be even higher and using a server to handle both control path and data path is reasonable and flexible in the future. An advantage is its support of dynamic load management. When a gateway is not busy, the server can run other functions in a SDN.

V. Communication Processor Trends

5.1 System requirement in the future

Faster and higher quality of service is always needed. Today, these demands push the emerging of new technologies, such as cloud services and 5G networks. However, for the traditional network industry, it is of great challenge to build new generation networking systems without substantial financial investments. The core structure of telecommunications infrastructure has to be more cost-efficient. To address the challenge, Software-defined networking (SDN) [45] and Network function virtualization (NFV) [46] have been introduced as new network solutions.

SDN is an architectural approach to decouple the network control from forwarding. It is directly programmable, enabling the infrastructure to be abstracted for applications. SDN deployment is expected to benefit the networking industry by increasing the functionality of the network while reducing costs and hardware complexity. NFV is the concept to transfer network functions from dedicated hardware appliances to software-based applications running on commercial off-the-shelf (COTS) equipment [47]. Major network equipment vendors have announced supports for NFV.

Much effort for the deployment of SDN and NFV has been made in the upper layers. However, in the lower infrastructure layer, hardware performance and efficiency for packet processing is still a fundamental challenge. In [48] it is demonstrated that only a hybrid approach with programmability/performance trade-off, e.g., a platform based on custom devices, NPs and server GPPs (general purpose processors) will provide an effective technology solution for SDN. For NFV Infrastructure, it is also required high computing capability through accelerations to meet the performance goals of virtualized functions.

5.2 The trends of "serverlization"

Along with the evolution of SDN and NFV, server, especially communication server, will play an essential role in network infrastructure devices. In the SDN architecture, SDN controllers are the kernel part, taking advantage of server virtualizations to support enormous protocols and separating network control from traditional network equipment [45]. In addition, NFV framework will consolidate various network equipments onto normal servers [49]. In other words, servers will replace traditional carrier-grade equipment dedicated for communication infrastructure.

There are mainly two popular multiprocessing architectures for servers: the SMP architecture, and HSA.

5.2.1 Classical architecture: SMP

SMP architecture has been the traditional parallel computing architecture for general purpose processors in x86/x64 servers and PCs. Nowadays, it remains the most popular in the high performance computing world, e.g., Intel Xeon and IBM Power.

In SMP, main memory is shared by homogeneous processors. The processors are interconnected through buses, crossbar switches, or on-chip networks. Each processor usually has a high-speed cache memory to speed up data access from main memory.

Application programs running on a SMP platform is often multithreaded to maximally use the multiple processor cores. OpenMP is a parallel programming model for shared memory multiprocessing system. It uses the fork-join model to execute programs in parallel.

The advantage of SMP is the homogeneity. All processors are based on one programming tool-chain using one instruction set. Tasks can be distributed and compiled in an easy way. All codes can be backward compatible and its early eco-environment is kept.

To improve the system performance of SMP platform, designers keeps adding more cores on the same chip. By increasing core numbers, the speedup becomes saturated according to the Amdahl's law.

5.2.2 A new trend for server: HSA

Applications and algorithms are different. Much higher performance can be achievable when different architectures are introduced to adapt to different algorithms for suitable applications.

However, heterogeneous architectures suffer from the poor support of parallel programming. Vendors have to develop their own programming tools, language extensions, and intrinsic kernels to match their specific hardware and assembly instruction-set features. To narrow the gap, parallel programming standards across heterogeneous platforms have been proposed, including OpenCL from Khronos Group, and DirectCompute from Microsoft. Despite all these efforts, programming cost has still been much higher comparing to that of SMP due to separated memory space, non-virtualized hardware, and so on.

HSA [50], being developed by the HSA Foundation since 2012, can relax the toughness. On the one hand, HSA provides architecture and coding environment to lower heterogeneous programmability barrier, and to reduce communication latency among computing elements. A unified programming model is provided, which allows programmers to write parallel applications that exploit both data-level parallelism and task-level parallelism. Several prominent features of HSA are developed for memory coherence control, task queuing, and context switching [51].

On the other hand, which could be more important, by following the hardware design and coding guides, developed codes could be shared on the HSAIL (HSA intermediate language) compiling level. The members of HSA can thus share coding kernel libraries to reduce the eco-system development cost on each side.

HSA development is supported by AMD, ARM, Imagination Technologies, MediaTek, Qualcomm, Samsung and Texas Instruments. These HSA members are strong experts in communications and consumer electronics. It might support code sharing and enhance eco-systems together.

Server processors incorporate with heterogeneous accelerators, even though they are not yet with HSA, for, example, Intel SMP with OPI (On Package Interface), Power 8 with CAPI of IBM, and APU of AMD.

On university research side, it will be significant to offer quality algorithm kernels using heterogeneous server processor instruction sets for communication applications.

5.2.3 Limitations of servers

In spite of the high performance and low cost provided by existing normal servers, they will not be able to handle foreseeable intensive and extensive- computation, and massive resource allocation requests. As a consequence, ETSI ISG (Industry Specification Group) presents schemes concerning accelerations for these puzzles, such as hardware accelerator, new Instruction Set Architecture (ISA), ASIP (e.g., BP, NP ZIP, and encryption processor) for heterogeneous accelerations [52]. Heterogeneous accelerations combining ASICs, DSPs, GPUs, FPGAs, or other accelerators, could be the most promising option in terms of R&D costs and SDN/NFV technological promotion.

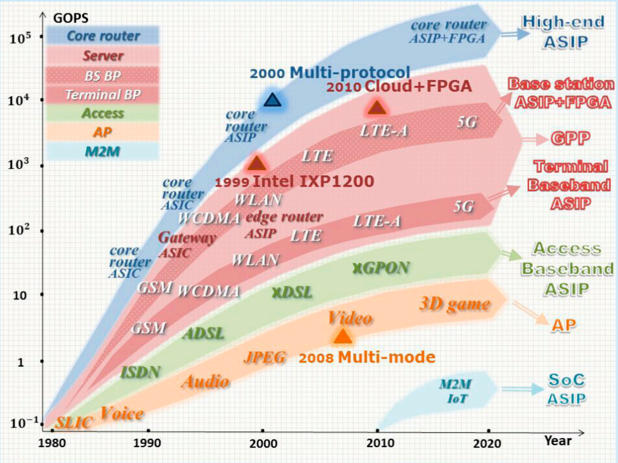


Fig.7 *The trends of communication processors*

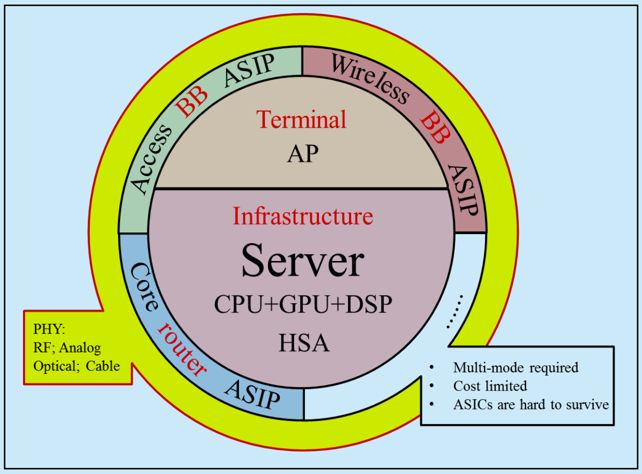


Fig.8 *The trends of communications system*

5.3 The trends of ASIP

Generally, performance over power consumption and performance over silicon cost are always the targets required by OEM of communication systems. ASIC seems the best solution. However, due to two essential problems, ASIC has been less popular. One reason is the high NRE (No Return Engineering) cost. The NRE cost includes at least the design cost, the full wafer silicon mask cost, and the silicon IP cost. The cost of an advanced SoC IC in a communication system can be tens of millions US Dollars. Another reason is the short TIM (Time In Market or life time in market) of ASIC.

The trade-off on performance and cost is actually the trade-off of flexibility. Very high flexibility can be offered by general processors (x86 or ARM). However, the performance over power or silicon is too low to be acceptable. On engineering side, flexibility is actually the compatibilities between neighboring standards (such as WCDMA and LTE) within fixed time (for example, five years are the sufficient TIM). The ASIP solution will thus be the best to reach the requirements on performance, power consumption, silicon cost, and suitable TIM.

ASIP is therefore a popular solution allocated in most positions in communication systems. On the core router side, ASIP dominates in the fast path. On the baseband side, ASIP has been the essential solution for smartphone and high-end xDSL. Inside a radio base station, ASIP is the best solution for programming easy algorithms. Likewise, inside AP and servers, ASIP has been used as engines for accelerations of Java, encryption, zip, and coding in communications.

To contribute ASIP for function acceleration will be even more meaningful. Design ASIP, its coding tools, and design methods for communications will be much needed. It is important to note that ASIP, in contrast to general purpose processor, is a broad name or family for any type of processor that is used to specific applications and fields, e.g., digital baseband processor, network processor and other kinds of DSPs.

5.4 Summary of trends

We summarize the trends of communication processor to 2020 in Figure 7. ASIP and servers are obviously driving the performance and reducing the cost. No doubt, server is the driver of flexibility. FPGA, which is a good hardware complement for performance and flexibility, will also play a certain role in high-end equipment.

VI. Conclusions

Requirements on performance over power consumption and over silicon cost keep increasing all the time. The pace is higher than the development of silicon technology. In addition, flexibility has become essential for multi-mode in more communication devices. Traditional ASIC is less used and ASIP is more important, as illustrated in Figure 8.

Serverlization is the trend of communication infrastructure. Server performance still needs a substantial improvement to narrow the gap between normal server and dedicated carrier-grade equipment. To ease this problem, research proposals are listed.

1. HSA in server: Superscalar silicon efficiency is very low, according to Amdahl's law, adding excessive homogeneous superscalar cores is meaningless, when the speedup is saturated. Accelerations from data parallel structures and application specific structures are of essential. Thus, there is a need to offer master-slave architecture and optional processors (GPUs, DSPs, CPUs, NPs, and accelerator ASIP) to featured tasks. Enough attention should thus be given to HSA. The programming on HSA is thus tougher. Coding-efficient architecture and programming methods shall be developed for accelerations. On SoC level, more HSA programming research for communication tasks should be offered, such as efficient task kernel design, cache coherent and task partition & dispatch to adapt communication task features.
2. Acceleration of programming hard algorithms using FPGA: We get very low performance when running non-parallelizable algorithms, such as bit manipulations, sorting, searching, FSM (Finite state machine, e.g. Huffman CODEC, CRC etc.). FPGA is so far also the best solution to accelerate algorithms with fixed long execution chain. In a communication system, much programming hard algorithms shall be accelerated. We thus need FPGA accelerations for communication tasks in servers. Challenges include tools for dynamic algorithm synthesis and efficient mapping on FPGA. Developing related algorithm kernels suitable for FPGA is also essential.
3. Some Algorithms, requiring excessive server computing power yet predictable, can be speeded -up by ASIP superior to FPGA. In recent years, ASIP, such as Virtual Java accelerator, encryption engine, zip engine, and network processor can be identified in servers following HSA programming. Servers for communications need more

ASIP and there is much space to improve qualities of these ASIP. Design of kernels running on ASIP shall receive considerable attention.

ACKNOWLEDGEMENT

The National High-Tech Research and Development Program of China (863 Program) 2014AA01A705.

References

[1] J. A. Harr, F. F. Taylor, and W. Ulrich, "Organization of No. 1 ESS Central Processor", *Bell Syst. Tech. J.*, vol. 43, no. 5, pp. 1845–1922, Sep. 1964.

[2] M. J. Kelly, "Introduction to PCM switching", *Auto Elec Tech*, vol. 12, p. 234, 1971.

[3] P. Marwedel, "The MIMOLA Design System: Detailed Description of the Software System," in *16th Design Automation Conference*, no. 1, pp. 59–63,1979.

[4] BDTi, "Selecting Application Processors for Mobile Multimedia," 2004.

[5] Carter L. Horney, "Cellphone Core Chip Trends : Market Analysis of Baseband , Application Processor , RF and Power Management Chips", 2015.

[6] D. Liu, "Baseband ASIP design for SDR", *China Commun.*, vol. 12, no. 7, pp. 60–72, Jul. 2015.

[7] American National Standards Institute, "ANSI T1.413-1998 'Network and Customer Installation Interfaces – Asymmetric Digital Subscriber Line (ADSL) Metallic Interface'", 1998.

[8] IEEE Communications Society,"{IEEE} Std 1901-2010 for Broadband over Power Line Networks: Medium Access Control and Physical Layer Specifications", *IEEE Std 1901-2010*, Dec. 2010.

[9] Yin-Tsung Hwang and Shin-Wen Chen, "Hardware efficient design for narrow band power line communication modem", in *TENCON 2011 - 2011 IEEE Region 10 Conference* , pp. 508–512, 2011.

[10] "G.984.1:Gigabit-capable passive optical networks (GPON)", [Online]. Available: http://www.itu.int/rec/T-REC-G.984.1/en.

[11] A. E. Kamal and B. F. Blietz, "Priority mechanism for the IEEE 802.3ah EPON", *IEEE Int. Conf. Commun. 2005*, vol. 3, no. C, pp. 1879–1883, 2005.

[12] U. Ramacher, "Software-Defined Radio Prospects for Multistandard Mobile Phones", *Computer (Long. Beach. Calif).*, vol. 40, no. 10, pp. 62–69, Oct. 2007.

[13] T. Suzuki, H. Yamada, T. Yamagishi, et al., "High-throughput, low-power software-defined radio using reconfigurable processors", *IEEE Micro*, vol. 31, pp. 19–28, 2011.

[14] D. Liu, A. Nilsson, E. Tell, D. Wu, and J. Eilert, "Bridging dream and reality: Programmable baseband processors for software-defined radio", *IEEE Commun. Mag.*, vol. 47, no. 9, pp. 134–140, Sep. 2009.

[15] V. Surducan, M. Moudgill, G. Nacer, et al., "The Sandblaster Software-Defined Radio Platform for Mobile 4G Wireless Communications", *Int. J. Digit. Multimed. Broadcast.*, vol. 2009, pp. 1–9, 2009.

[16] M. Woh, Y. Lin, S. Seo, et al., "From SODA to scotch: The evolution of a wireless baseband processor", in *2008 41st IEEE/ACM International Symposium on Microarchitecture*, pp. 152–163, 2008.

[17] T. Limberg, M. Winter, M. Bimberg, et al., "A Heterogeneous MPSoC with Hardware Supported Dynamic Task Scheduling for Software Defined Radio", *Design Automation Conference*, 2009.

[18] D. Liu, "Embedded DSP Processor Design", *Morgan Kaufmann*, 2008.

[19] A. Ghosh, N. Mangalvedhe, et al., "Heterogeneous cellular networks: From theory to practice", *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 54–64, Jun. 2012.

[20] "Small Cell SoC For Enterprise Markets." [Online]. Available: http://www.broadcom.com/products/broadband-access-and-modems/broadband-carrier-access/bcm61755.

[21] "Multicore DSP+ARM KeyStone II System-on-Chip (SoC)", 2013. [Online]. Available: http://www.ti.com/lit/ds/symlink/66ak2h14.pdf.

[22] "Delivering on the 1000x Small Cell Challenge – FSM99xx", [Online]. Available: https://www.qualcomm.com/news/onq/2013/06/04/delivering-1000x-small-cell-challenge-fsm99xx.

[23] "ITRS 2013 report.", 2013. [Online]. Available: http://public.itrs.net/reports.html..

[24] C. H. van Berkel, "Multi-core for mobile phones", in *2009 Design, Automation & Test in Europe Conference & Exhibition*, pp. 1260–1265, 2009.

[25] B. Noethen, O. Arnold, E. P. Adeva, et al., "10.7 A 105GOPS 36mm2 heterogeneous SDR MPSoC with energy-aware dynamic scheduling and iterative detection-decoding for 4G in 65nm CMOS", *Dig. Tech. Pap. - IEEE Int. Solid-State Circuits Conf.*, vol. 57, pp. 188–189, 2014.

[26] Z. Ziyuan, T. Shan, S. Yongtao, et al., "A 100 GOPS ASP based baseband processor for wireless communication", pp. 121–124, 2013.

[27] R. Hameed, W. Qadeer, M. Wachs, et al., "Understanding sources of inefficiency in general-purpose chips", *Proc. 37th Annu. Int. Symp. Comput. Archit. - ISCA '10*, p. 37, 2010.

[28] R. Giladi, "Network processors: architecture, programming, and implementation", *Morgan Kaufmann*, 2008.

[29] H. J. Chao, "Next generation routers", *Proc. IEEE*, vol. 90, no. 9, pp. 1518–1558, Sep. 2002.

[30] K. G. Coffman and a M. Odlyzko, "Internet growth : Is there a ' Moore ' s Law ' for data traffic ?", in *Handbook of massive data sets*, Springer US, pp. 47–93, 2002.

[31] D. S. F. D. Liu C, "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification", 1998.

[32] J. R. Allen, B. M. Bass, et al., "IBM PowerNP

network processor: Hardware, software, and applications", *IBM J. Res. Dev.*, vol. 47, no. 2.3, pp. 177–193, Mar. 2003.

[33] Intel corp., "Intel IXP 1200 Network Processor", 2001.

[34] T. Eklund, "The World's First 40Gbps (OC-768) Network Processor", *Presentation, Network Processor Forum*. 2001.

[35] Agere, "PayloadPlus Routing Switch Processor", *Lucent Technologies, Microelectronics Group*, 2000.

[36] S. McMahan, B. Erickson, et al., "A 600 MHz NT3 network processor", in *2003 IEEE International Solid-State Circuits Conference, 2003. Digest of Technical Papers. ISSCC.*, vol. 1, pp. 256–257, 2003.

[37] D. Husak, "Network Processors: A Definition and Comparison", *White paper, C-PORT*, 2000. [Online]. Available: http://www.cportcorp.com/solutions/docs/netprocessor_wp5-00.pdf.

[38] EZchip Technologies Ltd., "NP-1 10-Gigabit 7-Layer Network Processor", [Online]. Available: http://www.ezchip.com/html/pr_np-1.html.

[39] "IEEE P802.3bs 400GbE Adopted Timeline", 2014.

[40] "Sahasra^{TM} Processor", [Online]. Available: https://www.broadcom.com/products/Knowledge-Based-Processors/Layers-2---4/Sahasra-51500.

[41] "The Cisco Flow Processor: Cisco's Next Generation Network Processor Solution Overview", [Online]. Available: http://www.cisco.com/c/en/us/products/collateral/routers/asr-1000-series-aggregation-services-routers/solution_overview_c22-448936.html.

[42] "Industrial Communication Networks Fieldbus specifications-Part 3-12: Data-Link Layer Service Definition-Part 4-12: Data-link layer protocol specification-Type 12 elements", in *International Electrotechnical Commission*, 2007.

[43] http://www.arm.com/

[44] F. Koushanfar, V. Prabhu, M. Potkonjak, and J. M. Rabaey, "Processors for mobile applications", *Proc. 2000 Int. Conf. Comput. Des.*, pp. 603–608, 2000.

[45] Open Networking Foundation, "Software-Defined Networking: The New Norm for Networks [white paper]", *ONF White Pap.*, pp. 1–12, 2012.

[46] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, "NFV: state of the art, challenges, and implementation in next generation mobile networks (vEPC)", *IEEE Netw.*, vol. 28, no. 6, pp. 18–26, Nov. 2014.

[47] ESTI NFV INDUSTRY SPECIFICATION GROUP, "Network Functions Virtualisation (NFV); Infrastructure Overview", 2015. [Online]. Available: http://www.etsi.org/technologies-clusters/technologies/nfv.

[48] S. Sezer, S. Scott-Hayward, P. Chouhan, et al., "Are we ready for SDN? Implementation challenges for software-defined networks", *IEEE Commun. Mag.*, vol. 51, no. 7, pp. 36–43, Jul. 2013.

[49] ESTI NFV INDUSTRY SPECIFICATION GROUP, "Network Functions Virtualisation (NFV) White Paper(3)", 2015.

[50] G. Kyriazis, "Heterogeneous System Architecture : A Technical Review", pp. 1–18, 2012.

[51] HSA FOUNDATION, "HSA Platform System Architecture Specification," 2015. [Online]. Available: http://www.hsafoundation.com/html/HSA_Library.htm.

[52] ETSI NFV Industry Specification Group, "Virtualisation (NFV): Infrastructure; Compute Domain", 2014.

[53] https://www.broadcom.com/

Biographies

Liu Dake, IEEE Senior, is Professor and Director of ASIP Lab, Beijing Institute of Technology, China, and also Professor of Computer Engineering Division at the Department of Electrical Engineering of Linkoping University, Sweden. He got technology doctor degree from Linkoping University Sweden in 1995. Dake published more than 150 papers on journals and international conferences and holds 5 US patents. Dake's research interests are high-performance low-power ASIP (application specific instruction set processors), integration of on-chip multi-processors for communications. Dake has experiences also in design of communication systems and Radio frequency CMOS integrated circuits. Dake Liu is the co-founder and CTO of FreehandDSP AB, Stockholm Sweden, and the co-founder of Coresonic AB, Linkoping Sweden. Coresonic was acquired by MediaTek, Currently he is the professor assigned by the China Recruitment Program of Global Experts (1000 plan).

CAI Zhaoyun, received his B.E degree from Beijing Institute of Technology, China, in 2011. He is currently pursuing his Ph.D. degree in Beijing Institute of Technology, with research interests in ASIP, parallel computing, and architectures of radio baseband. *The corresponding author.120301@bit.edu.cn

WANG Wei, received his M.S. degrees in Computational Mathematics from ChengDu university of technology in 2012. He is currently pursuing his Ph.D. degree in Beijing Institute of Technology, with research interests in 5G PHY technology, low power baseband processor design, and parallel computing.