神威 E 级原型机互连网络和消息机制

高剑刚 卢宏生 何王全 任秀江 陈淑平 斯添浩
周 舟 胡舒凯 于 康 魏 迪
(国家并行计算机工程技术研究中心 北京 100190)

摘 要 本文描述了神威 E 级原型机的互连网络和消息机制. 神威 E 级原型机是继神威蓝光、神威·太湖之光之后神威家族的第三代计算机. 该计算机作为一台 E 级计算机的原型机,峰值性能 3.13 PFlops,其最大的特色之一就是采用 28 Gbps 传输技术,设计开发了新一代的神威高阶路由器和神威高性能网络接口两款芯片,在传统胖树的基础上,设计了双轨泛树拓扑结构,定义实现了新颖的神威消息原语和消息库,实现了一种基于包级粒度动态切换的双轨乱序消息机制,通信性能比神威·太湖之光互连网络提升了 4 倍,为神威 E 级计算机互连网络的研制奠定了基础.

关键词 多轨网络;泛树;高阶路由器;路由算法;网络接口;消息引擎;消息库
中图法分类号 TP302 **DOI 号** 10.11897/SP.J.1016.2021.00222

The Interconnection Network and Message Machinasim of Sunway Exascale Prototype System

GAO Jian-Gang LU Hong-Sheng HE Wang-Quan REN Xiu-Jiang CHEN Shu-Ping SI Tian-Hao
ZHOU Zhou HU Shu-Kai YU Kang WEI Di
(*National Research Center of Parallel Computer Engineering and Technology, Beijing* 100190)

Abstract The high-performance interconnection network is one of the main components of the high-performance computing system. It is responsible for the connection of computing nodes, storage nodes, and I/O devices in the high-performance computing system, and is responsible for the communication of all nodes in the high-performance computing system. There are a large number of parallel applications in high-performance computing systems that need to exchange data between different nodes (between computing nodes, between computing nodes and IO nodes, between computing nodes and storage nodes). High requirements are put forward for the communication delay and bandwidth of high-performance interconnection networks. A large number of high-performance computing systems have adopted customized interconnection networks to meet application requirements. The customized interconnection network can well meet the design requirements of high performance computing system, and can optimize the design of network performance such as communication delay and communication bandwidth to better meet the various communication requirements of high-performance computing systems and improve communication performance, thereby improving the actual operating performance of parallel applications in high-performance computing systems. Interconnection network design is an important means to improve network communication performance. At the same time, the

收稿日期:2020-05-15;在线发布日期:2020-08-25. 本课题得到国家重点研发计划项目(2016YFB0200500)资助.**高剑刚**,硕士,正高级工程师,主要研究方向为计算机体系结构、高性能互连网络. E-mail: 13701512205@139.com. **卢宏生**(通信作者),硕士,正高级工程师,主要研究方向为计算机体系结构、高性能互连网络. E-mail: lu_hongsheng@126.com. **何王全**,博士,正高级工程师,主要研究方向为计算机体系结构、并行语言设计. **任秀江**,博士,工程师,主要研究方向为计算机体系结构、高性能互连网络. **陈淑平**,硕士,高级工程师,主要研究方向为计算机体系结构、互连网络软件. **斯添浩**,硕士,工程师,主要研究方向为计算机体系结构、高性能互连网络. **周 舟**,博士,工程师,主要研究方向为计算机体系结构、高性能互连网络. **胡舒凯**,硕士,工程师,主要研究方向为计算机体系结构、高性能互连网络. **于 康**,博士,工程师,主要研究方向为计算机体系结构. **魏 迪**,硕士,工程师,主要研究方向为计算机体系结构.

message mechanism has a huge influence on communication performance. Even under the same topology and router conditions, different message mechanisms will still cause huge differences in communication performance. The customized features of customized networks are largely reflected in the ability to customize various message mechanisms. Each customized network has its own message mechanism and defines its own message protocol to meet its own special communication needs. The high-performance interconnection network and message mechanism are studied on the purpose of independent control. The communication performance must match the fast developing computing capability on the road to exascale system. The worldwide top supercomputers mainly select Mellanox InfiniBand, Cray Aries, Intel Onmi-path, and employ the 25 Gbps transmission technique to implement their interconnection network. The networks of the top domestic supercomputer, such as "Sunway Taihu Light" and "Tianhe 2", are constructed based on 14 Gbps transmission. The interconnection network and message mechanism of the Sunway exascale prototype system are introduced in this paper. Sunway exascale prototype system is the third-generation supercomputer of Sunway supercomputer family, after Sunway Blue Light and Sunway Taihu Light. As a pre-research project for the exascale system, the peak performance of this system is up to 3.13 PFlops. The interconnection network of this system is constructed based on two innovative Sunway chips: the Sunway high-radix router chip and Sunway high-performance network interface chip, depending on the 28 Gbps transmission technique. Moreover, a generalized fat-tree network topology is developed; an out-of-order message mechanism with dynamic packet-interleaved transmission in two rails is implemented; the efficient Sunway message verbs and library are designed. The communication performance of the prototype system improves 4 times compared with Sunway Taihu Light, and it therefore makes the solid technology foundation for Sunway exascale system. Sunway exascale prototype system makes the break-through on the key technologies of 28 Gbps transmission, high-radix router, high-performance network interface, high-efficient and reliable network architecture. Furthermore, Sunway network chipset of new generation is designed, and the network of Sunway exascale prototype system is constructed. They all contribute to the design of the domestic exascale supercomputer. The research achieves the goal of innovative design of the exascale system by constructing the large-scale verification system, mastering the techniques of new interconnection network architecture, and testing based on domestic components and parts.

Keywords　multi-rail network; generalized fat-tree topology; high-radix router chip; routing arithmetic; network interface; message engine; message library

1　引　言

虽然目前 TOP500 排名第一的"富岳"超级计算机性能已经达到 513 PFlops[①]，但鉴于超级计算机在科学研究、经济与社会方面的巨大作用，人类对计算能力的追求永无止境，世界各国都在加紧 E 级计算机的研制开发.

根据报道，美国预计 2021 年到 2023 年将提供 Aurora、FRONTIER、EI-CAPITAN 三台 E 级计算机，Intel、Cray、IBM、AMD、NVIDIA 和 HPE 六大超算巨头全部参与，其中 Aurrora 计算机 2021 年完成研制，峰值性能 1EFlops，计划成为美国第一台 E 级计算机. 三台计算机的共同之处是都将使用 Cray 公司最新的 Shasta 架构，其核心是 Cray 公司的 Slingshot 互连技术，Cray 公司研制了 Rosetta 高阶路由器芯片，该芯片具有 64 个端口，每个端口配置 4 个通道，采用 56 Gbps PAM4 传输技术，单端口传输速率达到 224 Gbps. Slingshot 互连采用了 Cray 独创的 HPC(High Performance Computing)以太

① Supercomputer RUGAKU. https://www.top500.org/system/179807/

网协议，在具备标准以太网优点的同时，也可以获得HPC网络的技术优势①.

日本Fugaku计算机系统，继承了“京K”计算与网络紧耦合的6D Trous结构，采用富士通自主研发的处理器A64FX构建，该处理器包含48个ARM架构的专用计算核心、4个辅助核心等部件，以及6个TNI模块和1个20端口片上Switch模块. Fugaku计算机的ToFuD网络采用了28 Gbps传输技术，实现了对ToFu2网络的全面升级②.

作为HPC网络领域技术领导者之一的Mellanox公司，其InfiniBand互连网络在2019年11月的TOP500中占比26.8%，目前TOP500中排名第一的Summit计算机就使用了Mellanox公司InfiniBand EDR产品，还有另外2台使用了其最新的HDR互连，传输速率达到50 Gbps. Mellanox公司的ConnectX-6芯片，采用50 Gbps PAM4技术，可配置成40个端口，每个端口200 Gbps，也可以配置成80个端口，每个端口100 Gpbs③.

法国Bull公司在高性能计算领域拥有强大的实力，其开发的BXI eXascale Interconnect采用了自行研发的48端口Switch芯片，每个端口性能达到100 Gbps，支持胖树、扁平蝴蝶[1]、蜻蜓等多种结构，支持最大64 K节点[2].

为了实现从P级到E级的跨越，探索应对访存墙、通信墙、可靠性墙、能耗墙、可扩展性墙等挑战④的有效技术和策略，中国研制了3台E级原型机，分别是曙光E级原型机、天河E级原型机和神威E级原型机. 其中神威E级原型机由1024个申威26010+众核处理器（简称SW26010+）组成，每个处理器包含4个运算控制核心和256个运算核心. 2个处理器通过PCIe3.0连接到同一片网络接口芯片，构成一个运算节点，每个运算节点峰值运算能力6.12 TFlops. 512个运算节点通过神威互连网络相连，组成神威E级原型机. 该原型机峰值性能达到3.13 PFlops，位居神威·太湖之光、天河2号、天河1号之后，在2018年中国超算TOP100中排名第4⑤，已安装在国家济南超算中心.

神威互连网络是神威E级原型机的重要组成部分，由网络硬件和网络软件组成. 如图1所示，网络硬件主要由新一代神威路由器芯片和网络接口芯片构建而成. 网络软件包括网络驱动、消息库、网络虚拟化、MPI、TCP/IP和网络管理软件等. 本文将描述神威E级原型系统的互连网络，具体安排如下：第2节描述互连网络系统硬件，主要包括网络组成和拓扑结构、神威高阶路由器芯片及其关键技术、神威双端口网络接口芯片及其关键技术；第3节描述互连网络软件，重点描述消息库和网络虚拟化支持；第4节介绍神威互连网络的可靠性体系；第5节介绍原型机的基础性能；第6节对神威互连网络的特点进行总结.

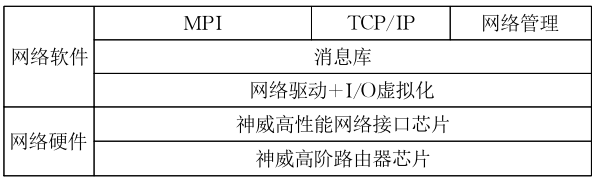

图1 神威互连网络系统

2 互连网络硬件

神威E级原型机互连网络由2台576端口交换机组成，每个运算节点通过一片网络接口芯片的2个端口分别连接到2台交换机. 交换机采用了leaf-spine结构[3]，其中第一层采用了32片神威路由器芯片（后文统一简称SWHRC：Sunway High-radix Router Chip），第二层采用了18片SWHRC芯片，每片SWHRC芯片40个端口，其中第一层芯片的18个端口用于连接光纤，18个端口用于连接第二层的18片SWHRC，具体如图2所示.

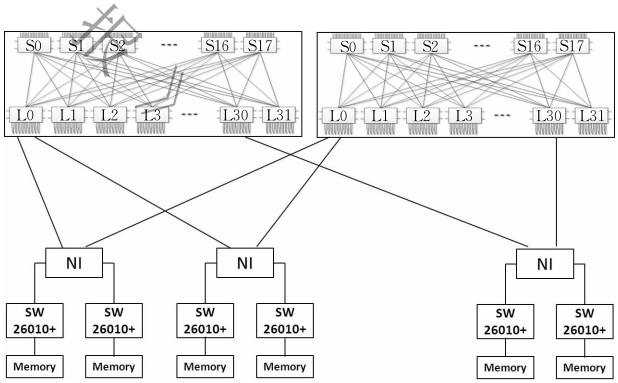

图2 神威E级原型机互连网络结构

① Inside Rosetta: The Engine Behind Cray's Slingshot Exascale-Era Interconnect. https://fuse.wikichip.org/news/3293/inside-rosetta-the-engine-behind-crays-slingshot-exascale-era-interconnect

② Yuichiro Ajima. https://www.fujitsu.com/global/Images/the-tofu-Interconnect-d-for-super computer-fugaku.pdf

③ ConnectX-6 Single/Dual-Port Adapter supporting 200 Gb/s with VPI. http://www.mellanox.com/products/infiniband-adapters/connectx-6

④ Steinmacher-Burow B, Gara A. Some challenges on road from petascale to exascale. http://www.physik.uniregensburg.de/forschung/wettig/workshops/APQ_April2010/talks/20100414%20IQCD%20RegensburgSteinmacher-Burowv07.pdf10.1093/nsr/nwu002.html

⑤ 2018年中国高性能计算机性能TOP100排行榜. http://www.hpc100.cn/top100/17

相比较神威·太湖之光实现 1024 个处理器互连使用了 4 层树网，神威 E 级原型机互连 1024 个处理器只使用了 2 层树网，网络直径减少了 4 步. 这主要得益于我们开发的新一代双端口网络接口芯片和双轨网络技术.

2.1　双轨泛树结构

胖树网络由于对各种通信模式都具有很好的适应性，同时又比多维环网有更短的网络直径，因此被很多高性能计算系统所采用. 但胖树网络存在一个众所周知的问题. 尽管胖树网络中的任意两点，比如 A 和 B，存在多条路径，但从 A 到 B 的上行路径一旦确定，则其下行路径就是唯一的，一旦该下行路径上出现故障，A 到 B 就无法正常通信. 为了克服胖树的这一缺点，提高胖树网络路径故障后的可达性，我们对标准胖树网络进行了改进. 定义 $R[i]$ 表示交换机第一层的第 i 号路由器，$i\in[0,N]$. 根据前面的描述，$R[i]$ 已经使用了 36 个端口，还剩余 4 个空余端口，将 $R[2m]$ 和 $R[2m+1]$ 剩余的 4 个端口互连（其中 m 是 $i/2$ 取整），$R[2m]$ 和 $R[2m+1]$ 互称为“兄弟”. 相对于标准胖树结构，我们称这种改进后的胖树为泛树，这对于提高超大规模系统网络的可靠性具有重要的意义.

如图 3 所示，在泛树结构中，当 A 到 B 的下行路径故障后，数据包下行时仍然可以通过先到达 B 所连接的开关的兄弟开关，再到 B 所连接的开关而顺利到达 B. 极限的情况下即使 B 所连接路由器的 18 个上行端口全部故障，其他节点仍然可以通过 B 所连接路由器的兄弟到达 B. 由于在网络中，两个互为兄弟的路由器芯片安装在同一块印制板上，因此实现兄弟路由器间的互连并未增加额外的开销. 事实上，泛树结构可以定义为在标准树网上增加同层路由器间互连的一种结构，该结构比标准树结构有更好的故障容错特性.

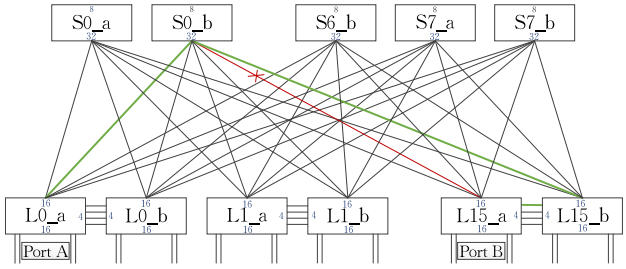

图 3　泛树故障路由示意图

2.2　SWHRC 高阶路由器芯片

SWHRC 芯片是构建神威互连网络的核心器件，实现了网络中物理层、链路层和网络层的功能. 如图 4 所示，SWHRC 芯片集成了 160 个速率 28Gbps

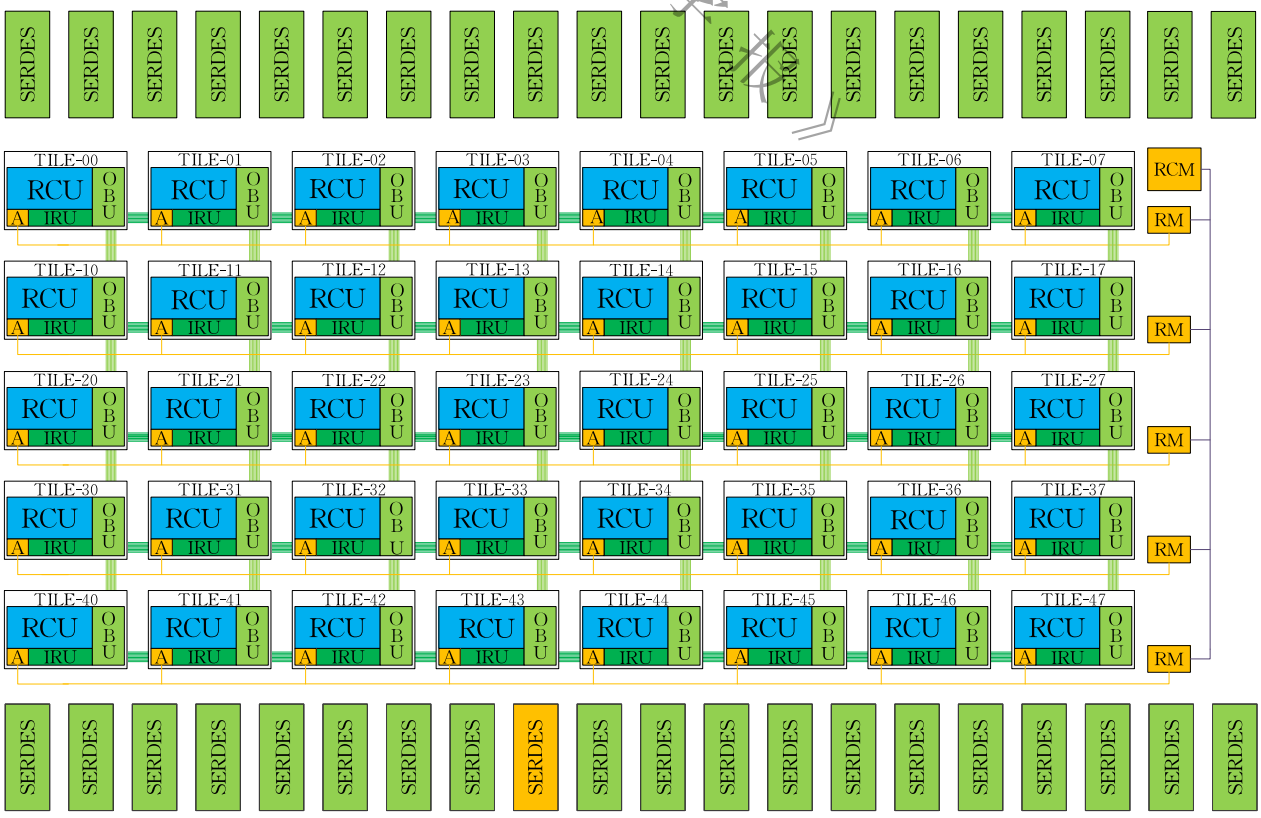

图 4　SWHRC 路由器芯片结构图

的 SerDes,芯片总吞吐率为 9 Tb/s. 芯片包含 40 个端口和 40 个交换子模块,每个交换子模块主要包含 IRU(Input and Router Unit)、RCU(Row Buffer and Crossbar Unit)、OBMU(Output Buffer Manager Unit)等部件,40 个交换子模块采用 8×5 阵列式排布.

(1)缓冲管理

每个端口包含四个虚通道. 输入缓冲采用动态共享机制,支持可配置的私有信用和共享信用,提高输入缓冲的利用率. 每个交换子模块内还包含 8 个行缓冲和 5 个列缓冲,分别用于接收同一行和同一列的数据. 这种缓冲管理机制和缓冲数量的设置既降低了芯片设计开销和实现难度,同时也降低了头阻塞效应,提升了数据吞吐率.

(2)流量控制

SWHRC 芯片采用 VCT(Virtual Cut Through)流控机制,为了避免故障包对网络带宽的浪费,增加对链路带宽变化的适应性,数据链路层支持 VCT 与 SAF(Save And Forward)两种机制的动态切换. SWHRC 芯片采用基于信用的流控,可以控制数据流按照信用多少提前发送消息包,提升数据的传输效率,降低延迟. 信用以数据块为单位,每个数据块包含 4 个数据切片,每个数据切片大小为 16 B. 拥塞预感知装置依据下游信用释放的速率回馈上游,从而使上游能更早感受到下游的拥塞情况,避免过度拥塞情况的出现.

(3)路由策略

为了保持消息的路径灵活度,并满足不同消息对路径选择的不同需求,SWHRC 芯片的路由策略支持源路由和分布式查表路由. 每个消息包携带路由策略选择标识,路由器根据消息包标识决定选择何种路由方式进行路径选择. 分布式查表路由支持确定性路由策略与自适应路由策略.

神威超级计算机包括神威蓝光、神威·太湖之光、神威 E 级原型机均特别重视应用的局部性特征和分区概念,在体系结构的设计中一直强调并坚持超节点的设计思想. 每个系统都可以认为是由一组超节点通过高性能骨干网互连构成. 基于上述两级网络结构,SWHRC 芯片实现了一种两级结构的路由表. 超节点内路由表和超节点外路由表. 每个接入神威互连网络的端口和每片 SWHRC 芯片都有一个唯一的 GUID(Global Unite ID)编号. GUID 包含超节点号和超节点内编号. 当目标 GUID 的超节点号和本 SWHRC 的超节点号相同时,就使用超节点内路由表路由,否则就使用超节点外路由表路由. 在实际的设计中两个表采用一个片上 RAM 实现,大小可配置. SWHRC 芯片每个端口均配置了一个两级结构路由表,支持系统最大节点规模为 256K,相比一级结构路由表存储器容量节省约 99.5%.

SWHRC 每个路由表项包含若干个目标端口,自适应路由决策模块根据下一级缓冲信用选择一个相对最空闲的端口做为输出;而确定性路由则根据条目内容顺序选择第一个可用的端口,结合拓扑特性,在部分链路故障时在修改路由表之前仍可以保证消息正确传输.

SWHRC 路由器支持点到点路由的同时,每个端口都有独立的 32 条多播路由表,支持 1 到 40 个端口的多播路由.

(4)维护管理

在维护管理方面,SWHRC 芯片同时实现了带外低速通路和带内高速通路的访问寄存器接口,带外管理通路基于标准 JTAG 协议;带内管理通路复用高速链路,实现自定义高速管理协议. 管理部件能同时处理带外和带内寄存器访问.

每个端口都有大量可配置寄存器和状态寄存器,包括高速链路配置和核心交换部件的配置、端口各虚通道接收/发送的包/flit 数量、缓冲占用、链路繁忙程度等. 针对阵列式交换架构,我们提出了一种片上两级并行管理架构,RCM(Router Control and Manage system)可同时作用五个 RM(Row Manage unit),每个 RM 可同时管理一行上的八个管理代理 MA(Manager Agent),实现一条管理命令同时作用于多个端口,极大地加快路由芯片的初始化速度与性能管理.

2.3 高性能网络接口芯片

神威高性能网络接口芯片(以下简称 SWHNI:Sunway High-performance Network Interface chip)是负责处理器间高速通信的核心器件,采用硬件方式实现了网络传输层功能,通过提供丰富的消息类型、灵活鲁棒的消息调度机制为处理器提供了高带宽、低延迟的数据传输能力. SWHNI 芯片提供了两个 PCIe 接口和两个网络端口. PCIe 接口支持 PCIe3.0 标准,每个 PCIe 接口包含 16 个通道,每个通道链路速率 8 Gbps. 每个网络端口包含 4 个通道,每个通道链路速率 28 Gbps. SWHNI 芯片最大数据吞吐率为 448 Gbps.

DESDP 结构

SWHNI 芯片设计了一种双引擎共享双端口的

结构(以下简称 DESDP:Double Engine Share Double Port),如图 5 所示,SWHNI 主要包括两个消息引擎和两个网络端口.消息引擎是消息机制的核心部件,负责消息的发送、接收,由 SU(Send Unit)、RU(Receive Unit)、PIU(PCIe Interface Unit)组成.NPU(Networt Port Unit)部件实现网络的物理层和链路层功能.两个消息引擎通过内部交叉开关可以共享两个网络端口.

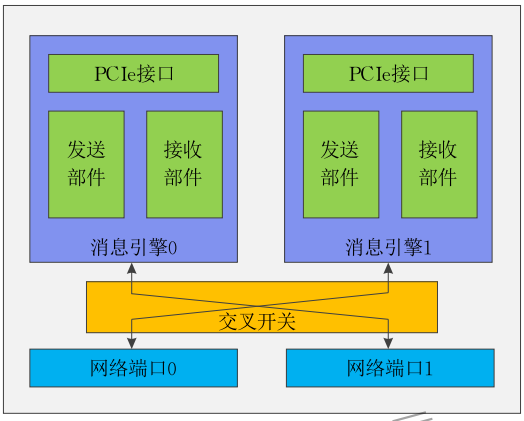

图 5　DESDP 结构

DESDP 架构使 SWHNI 具备以下三个优势:

(1)实现 SWHNI 高效支持两种节点连网模式.如图 6(a)所示,SWHNI 支持一个处理器通过 2 个 PCIe 接口和一片 SWHNI 相连,通过构建双轨网络模式提升处理器的通信能力.如图 6(b)所示,SWHNI 也支持同时连接两个不同的处理器,以降低网络系统规模与开销.

(2)实现 SWHNI 高效支持三种通信模式.模式一如图 6(a)所示,通过消息引擎内部通路可以实现使用同一 PCIe 接口的进程间通信.模式二如图 6(b)所示,通过片内交叉开关实现使用不同 PCIe 接口的进程间通信.模式三如图 6(c)所示,通过网络端口实现使用不同 SWHNI 芯片的进程间通信.多种通信模式由 SWHNI 芯片根据通信目标硬件自动判别,有助于编程的一致性且通信性能符合局部性特征.

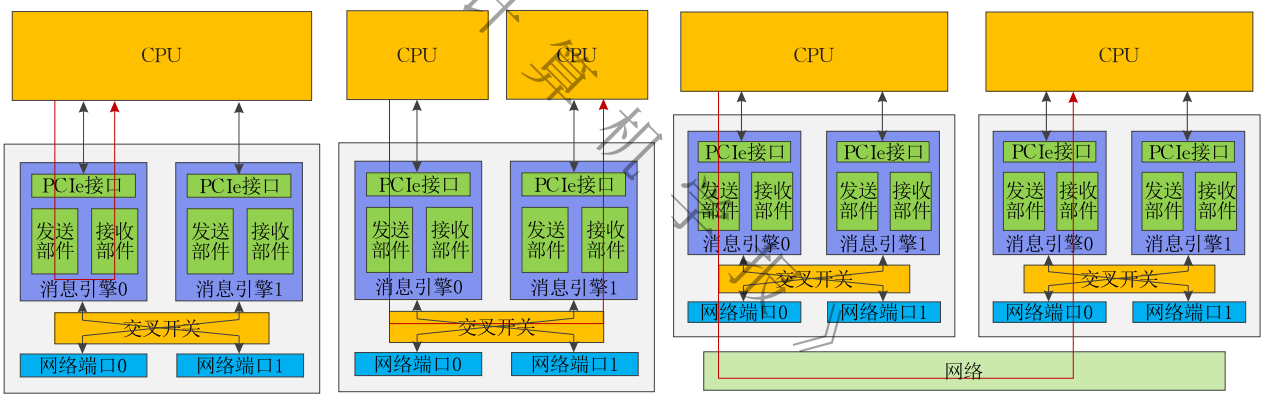

图 6　SWHNI 连接模式

(3)通过共享网络端口模式提高了数据吞吐率和通信可靠性.对于两个网络端口的管理,SWHNI 芯片实现了一种基于细粒度包级动态网络端口选择机制,每个数据包可以根据当时端口的链路好坏、忙闲情况动态地选择最终的上网端口,这种选择策略是基于神威互连网络实现的一种高效乱序的消息机制,在提升网络端口利用率的同时提高了通信可靠性.

乱序消息引擎

SWHNI 消息引擎是消息处理的核心部件,主要由发送部件 SU、接收部件 RU、PCIe 接口部件 PIU 组成,具有如下技术特色:

(1)支持多处理器多进程

SWHNI 消息引擎支持多处理器多进程资源共享,最多支持 128 个虚拟接口.每个虚拟接口以队列方式提供软件使用.各队列由软硬件协同管理,允许多进程独立或共享使用.发送队列作为消息发起的接口,支持多进程并行投递消息描述符,各队列采用均匀轮转方式仲裁,仲裁上的消息由消息引擎轮转调度执行,消息引擎根据消息长度和快速标记按优先级调度消息的执行.同一时刻一个引擎最多飞行的消息数为 256 个.多队列并行消息机制实现了多进程共享硬件资源,同时也有利于控制网络拥塞.

(2)支持旁路操作系统零拷贝

SWHNI 芯片通过地址映射的方式,支持软件直接通过用户层访问使用.SWHNI 芯片实现了远程 RDMA 功能,消息空间支持物理地址、页式虚地址和段式虚地址访问,其中页式虚地址支持最小 4 KB、

最大 8 MB 等多种页粒度. SWHNI 内置 2048 条目的 TLB,支持页表的预取. SWHNI 芯片实现片上地址代换和 DMA(Direct Memory Access)功能,消息数据的读写不需要处理器干预,允许用户直接访问消息空间数据,实现了旁路操作系统零拷贝数据传输.

(3)支持消息双轨网络乱序执行

SWHNI 芯片集成两个网络端口,实现消息引擎的双轨网络连接,消息引擎通过交叉开关与网络端口连接,共享两个网络端口. 消息引擎采用乱序提交的消息执行策略,即处理相同和不同发送队列的消息时,可以不控制消息间的执行顺序,尽可能地发挥硬件性能,实现消息间的快速调度;采用数据流驱动机制,实现基于乱序的数据流水传输. 消息包支持两种上网方式:指定上网端口和自适应上网端口,第一种方式允许用户在描述符中指定消息包的上网端口号,所有消息包从指定端口上网;第二种方式由硬件自动根据端口忙闲情况,动态分配空闲的端口上网,并且支持同一消息的不同消息包从不同端口上网,以达到双轨网络的流量均衡和通信带宽的最大化.

(4)支持软件定义的硬件集合通信

SWHNI 消息引擎实现了硬件集合通信机制,采用软件定义逻辑树与硬件链表融合的模式,实现了高效灵活的同步(Barrier)、多播(Bcast)、归约(Reduce)等集合消息,其中归约操作的数据类型支持单/双精度浮点数、32 bit/64 bit 定点数,计算类型包括累加、求最大/小值、按位与/或/异或等. 软件将集合通信拓扑结构定义写入到集合描述符中,硬件按照描述符中定义完成具体操作. 归约计算中对顺序有要求的浮点累加操作可以通过定义特殊树结构来控制数据计算过程. 所有集合消息执行流程统一为归集和广播两个过程,具体分为归集请求发送、逐层数据计算(归约操作)和广播结果处理. 其中归集过程由 SWHNI 芯片通过定制的可靠消息完成,保证归集通信可靠;计算过程在 SWHNI 上完全由硬件实现;广播则由根节点 SWHNI 芯片在完成全部叶子操作后自动发起,由 SWHRC 芯片按照预先配置的多播链完成数据复制和传输. 多播链由软件根据作业管理策略配置,支持粗粒度模式匹配. 即使目标节点接收到不属于自身的多播也会被目标方 SWHNI 根据进程信息丢弃. 该集合通信机制突破网络拓扑结构限制,增强了在不同网络拓扑下的适应性,并在简化硬件设计和软件使用模式的同时,解决了受限的集合树硬件资源和巨大的软件需求间的矛盾,消除了动态修改链表开销,大幅提高了集合通信的处理能力、可扩展性和实用性.

(5)支持快速作业退出

SWHNI 芯片的消息引擎实现了高效的用户作业退出机制,作业退出时,用户只需要注销作业相关的片上资源即可. 资源注销后片上残留消息和网上残留包由消息引擎自动处理,不需要用户干预. 残留消息的处理不影响其它用户作业和后续提交作业的正常运行,系统保证主存和片上空间数据不被污染. 该机制可实现残留消息的快速清理和片上资源的快速注销回收.

神威消息原语

SWHNI 芯片通过消息原语的方式实现高效的数据通信. 消息原语主要包括点到点消息、集合通信和快速短消息 FSM(Fast Short Message)三大类型.

(1)点对点消息

实现了 RDMA(Remote Direct Memory Access)、异步队列、管理等多种点对点通信类型,满足不同条件下、不同通信模式的需求.

RDMA 通信操作. 与很多主流网络[4-7]类似,SWHNI 芯片硬件实现了 RDMA,包括 RDMAW(Remote Direct Memory Access Write)和 RDMAR(Remote Direct Memory Access Read). RDMAW 将本地存储空间的数据写入到远程节点的存储空间. RDMAR 将远程存储空间的数据读到本地存储空间,采用把 RDMAR 在目标方转化成 R2W(Read message to Write message)的实现策略. 神威网络支持两种类型的 RDMAW,一种是通过消息事件通知数据传输完成,另一种是通过对远程回答字原子加 1 来通知数据传输完成,如图 7 所示,这种

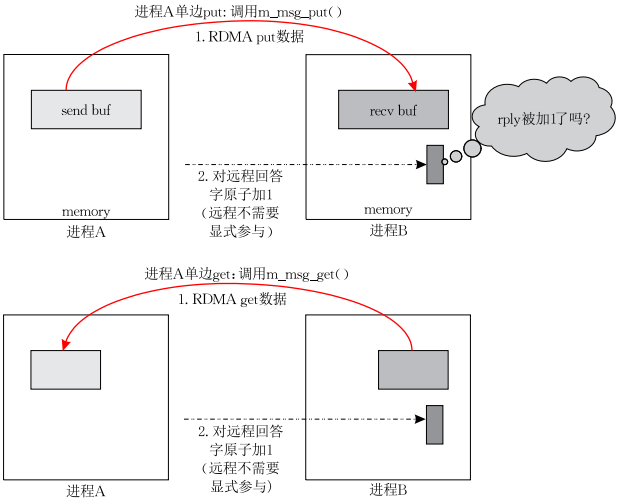

图 7 使用远程回答字的单边 put 和单边 get 消息

消息特别适合上层语言实现精简的单边消息机制，与前一种方式相比，消息被动方判断消息是否完成，只需要读一下回答字的内容是否更新，开销更小，使用也非常方便，这是神威消息协议的一个特色.

异步队列消息，消息数据直接写入指定的接收队列环，不需要源方指定目标地址，异步队列消息支持可靠和不可靠两种类型，以适应不同的应用场景.

管理消息支持源路由和目标路由两种方式，源路由管理消息用于拓扑探查、网络的初始化和更新路由表，目标路由管理消息可用于性能管理等. 设计了两级管理命令映射配置机制，为软件提供可配置的管理命令机制，方便灵活定制网络管理命令.

(2) 集合消息

集合消息(Collective Communication)：包含同步、多播、归约等消息类型，集合消息的流程为，请求发送、逐层计算(归约操作)和结果处理，集合消息由软件构建集合树结构并投递集合描述符，其余过程由消息引擎独立完成，软件干预程度低，相比点对点消息实现的集合操作延时更低.

(3) FSM 消息

快速短消息 FSM：消息的描述符和数据通过 PIO 直接写到片上储存体，消息引擎收到门铃后直接从片上读取消息描述符并解析执行，组包时直接从片上读取消息数据，从而减少了 2 次跨 PCIe 的访存开销，极大地降低了消息延迟. FSM 消息支持对远程节点主存空间和片上 IO 空间的读、写操作，具有较高的使用灵活性.

3　消息软件

神威 E 级原型机网络软件由 SWHNI 驱动、神威消息库及网络管理软件组成，如图 8 所示. SWHNI 驱动位于内核层，负责初始化、配置、管理网络接口芯片中的各类通信资源，并提供虚拟化、兼容 TCP/IP 支持等. 神威消息库位于驱动层之上，为 MPI 等上层用户提供高带宽、低延迟的消息通信编程接口. 网络管理软件位于最高层，用于实现网络系统的初始化与配置、网络事件处理、网络故障诊断等功能. 本节重点介绍消息软件的核心部分消息库和虚拟化支撑.

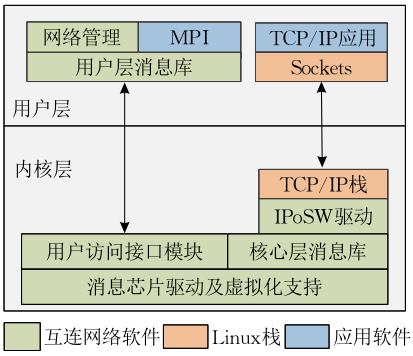

图 8　神威网络软件组成

3.1　神威消息库

神威消息库基于神威消息原语开发，主要负责向上层并行语言、文件系统等提供通信编程接口，包括为上层应用提供资源分配及管理接口、消息收发接口，为其提供高效可靠的数据传输能力，提供透明的容错机制，处理各类异常. 我们自主定义了 SWverbs 编程接口，包括设备操作 API、内存句柄注册 API、队列操作 API、工作请求 API、完成及事件通知 API 等类别，共 30 余个 API. SWverbs 在支持一些常见功能的同时，充分体现了对神威网络提供的双轨机制、RDMA、FSM 消息等新机制的支持，主要功能和技术特色如下：

(1) 支持异步队列消息、RDMA Write、RDMA Read 等传统的消息类型，提供对发送队列 SQ(Send Queue)、接收队列 RQ(Receive Queue)、完成队列 CQ(Complete Queue)等通信资源的管理，支持为用户空间内存注册句柄，提供发送及接收请求的异步查询方式以支持计算与通信并行，支持完成事件通知等.

(2) 支持硬件广播、归约等集合操作类型，相比软件方式实现的广播、归约操作，有了明显的性能提升.

(3) 针对高性能计算领域的典型通信模式，提出了多种新消息类型，包括带回答字的 RDMAW 消息、带通知的 RDMAW 消息、FSM 快速短消息. 传统的 RDMAW 消息是单边消息，发送方发起 RDMAW 操作后，不需要消息接收方的参与，消息完成后接收方不会收到任何通知. 因此发送方在发起 RDMAW 后，需要再发送一个异步队列消息到被动方，以通知 RDMAW 消息的完成，这会产生不小的性能开销，对延迟敏感型应用会产生较大影响. 神威消息库对此进行了改进，提供了 2 种新的 RDMA 通信操作：

一是带通知的 RDMAW，相当于将传统的 RDMAW 与异步队列消息合并到一条消息中进行传送，接收方在收到异步队列消息后，即标志着对应的 RDMAW 消息已完成，从而可以节省发送方投递异步队列及查询完成的时间开销.

二是带回答字的 RDMAW，当 RDMAW 完成后，在接收方会将用户指定的计数器加 1，以通知用户 RDMAW 的完成个数，从而可以减少一次异步队列的发送开销，提升 RDMA 操作的性能.

除此之外，还提供对 FSM 消息的支持. FSM 消息采用专用路径发送消息，可获得比普通消息更小的延迟，最大可支持 128 字节长度的消息，长度较小的控制消息可通过 FSM 快速消息进行发送.

(4)从优化通信模式、减少数据拷贝等方面对消息库关键路径进行了优化，为上层用户提供低延迟、高带宽的网络传输能力. 主要手段包括，精心设计发送描述符格式，压缩到 64 B 以内；针对 FSM 消息，支持将多个缓冲区的数据组织成一条 FSM 发送；支持虚拟 RQ 和 Raw RQ 两种方式，满足不同应用场景的需求.

(5)提供高扩展可靠数据报传输服务. 在大规模并行应用课题中，每个进程需要同时与数十万个其它进程进行通信. 传统的消息库中如 ibverbs 提供 RC 可靠连接及 UD 不可靠数据报两种传输服务，不能满足 E 级系统的需求. 例如，随着系统规模不断扩大，采用可靠连接服务实现全连接通信时，需要创建大量的连接，每个连接都需要分配独立的 QP 通信资源及内存资源，从而带来严重的可扩展性问题. 而采用不可靠数据报服务时，消息库不保证数据的可靠传输. 在消息库层级上提供可靠数据报服务是解决上述问题的一种重要途径. 一方面，一个进程跟多个进程通信时，仅需要创建一个发送队列、一个接收队列即可，在发送数据时，所有的请求都投递到同一个发送队列中；而在接收数据时，所有的接收请求也投递到同一个接收队列，使得应用程序在各个节点间建立全连接时仅需分配一个发送队列、一个接收队列，大大降低了建立全连接时所需的通信资源及内存资源数量，有利于提高系统的可扩展性. 另一方面，数据报的可靠传输由消息库保证，对上层用户透明，使得上层用户不用关心底层的可靠传输实现机制.

3.2 设备虚拟化

无论是在云计算还是在 NFV 中，虚拟机都是重要的支撑技术. 神威消息软件为 SWHNI 芯片提供了虚拟化支持，为每个虚拟机提供了虚拟的消息接口及高带宽低延迟的通信服务. 目前基于 PCIe 总线的 IO 设备虚拟化通用标准是 SRIOV，但 SRIOV 需要网卡设备以硬件或固件的形式提供支持，硬件设计复杂度很高. 我们采用软硬件协同的方式实现了网卡虚拟化功能，支持硬件资源的动态、按需分配，支持容器式虚拟机快速部署和故障物理隔离，而且虚拟机网络性能损失极小.

传统的虚拟机资源分配技术将通信资源静态分配给不同的虚拟机，容易造成硬件资源的浪费. 与传统的虚拟机资源分配技术不同，我们采用宿主机代理的硬件资源动态分配技术(如图 9 所示)，将硬件

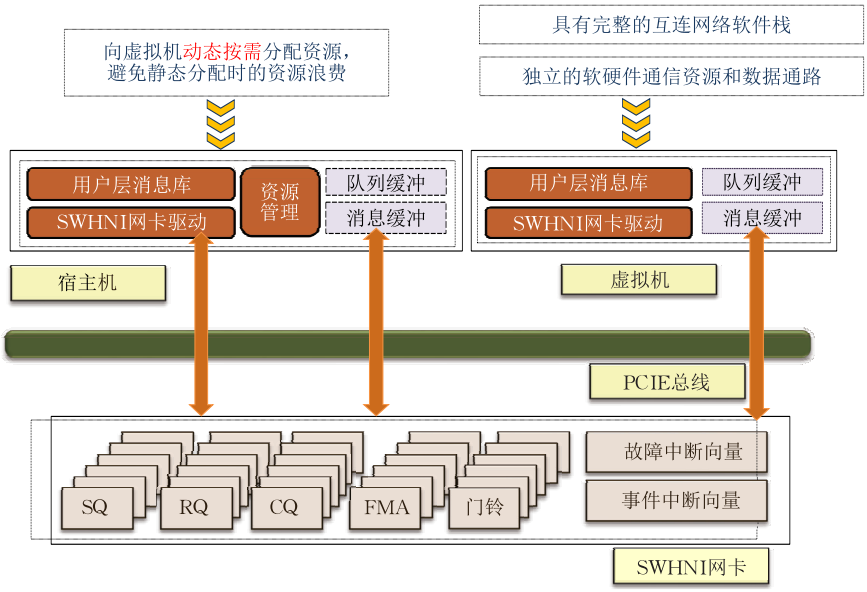

图 9 SWHNI 虚拟化

通信资源按需动态地分配给不同虚拟机使用，从而避免静态绑定带来的资源浪费问题. 另外，每个虚拟机完成资源分配工作后，可以独立工作，具有良好的物理隔离性，以支持 SWHNI 芯片驱动及消息库的容器化快速部署.

在消息通信方面，采用虚拟机消息旁路核心技术，使各虚拟机上具有完整的互连网络软件栈、独立的软硬件通信资源和数据通路，从而使虚拟机的旁路虚拟机操作系统及宿主机操作系统可以直接访问硬件通信资源，保证虚拟机上消息性能与宿主机相比没有显著下降.

4　网络可靠性

网络规模的扩大和传输速率的提升使互连网络的可靠性面临了新的巨大挑战. 应对挑战，新一代神威互连网络建立了从物理层到传输层的层次式可靠性保障体系. 物理层，高速串行传输技术是集成电路封装引脚受限条件下提升通信链路峰值带宽的最有效手段. 神威互连网络率先在国际超级计算机研制中实现 28 Gbps 高速传输，通过各种链路均衡结合阻抗控制、优化的 PCB 盲孔设计等技术，实现 28 Gbps 传输信号在插入损耗 30 dB 以上时误码率小于 10^{-12}，高于业界标准 6 个数量级，有力保障了包含数百万条链路的超级计算机的基础可靠性；采用 RS(528,514) 编码实现前向纠错功能，可以实时纠正连续 70 位数据错或 7 个 10 位的符号错，进一步提升了超大规模互连网络的可靠性. 针对链路随机错，实现了基于 CRC32 编码校验的链路层重传机制；针对部分链路固定故障，支持链路宽度从 4 个通道自动协商到 2 个通道、1 个通道. 网络层利用双轨泛树网络结构的冗余路径，结合自适应路由策略、路由重构策略和传输层消息重传技术，在单端口失效、单芯片失效、甚至一台完整的交换机失效情况下，仍然能够保持应用程序不间断消息服务.

采用软硬件协同的容错机制，解决异步队列消息、RDMA、集合三类消息的可靠传输问题.

(1) 在异步队列消息容错方面，采用基于轻量级滑动窗口的异步队列消息软硬件协同可靠传输技术，采用多个并发的轻量级滑动窗口，通过发送方消息重传及接收方消息排重机制，由软硬件协作共同实现异步消息的可靠传输，为上层用户提供透明的异步队列可靠传输服务，以解决异步队列消息的高效可靠传输问题. 与传统滑动窗口相比，通信双方无需进行握手和信用反馈，无需对乱序数据包进行缓存，维护并发窗口的内存开销和 CPU 开销都非常小.

(2) 在 RDMA 容错方面，在软件层采用基于消息应答协议的可靠机制，结合对硬件资源的状态探查，综合判断消息未正常完成的状态，通过重发数据、重发完成通知等操作，保证数据可靠传输，在消息库实现了 RDMA 数据可靠传输，带宽接近硬件峰值.

(3) 在集合消息容错方面，通过软件超时发现集合通信错误，软件定义逻辑通信树，为叶节点、父节点以及根节点分别设计针对性的探测和应答策略，结合对硬件寄存器、消息缓冲区、消息队列的内容进行读取，帮助通信过程中各节点确认全局的集合通信状态，对出现的错误进行识别和判断，实现精准容错决策和全局容错方法.

神威互连网络系统建立了一体化层次式容错体系. 硬件在软件无感知情况下解决了系统中发生的 99.9%以上错误，且对性能几乎没有任何影响. 软硬协同容错机制进一步增强了容错体系的完备性，有效保障整机系统可用性在 99.9%以上.

5　性　能

5.1　基础性能

神威 E 级原型机互连网络采用了新一代神威网络芯片组研制，基础通信能力有了明显提升. 表 1 对神威 E 级原型机网络基础性能与 2019 年 11 月份 TOP500 榜单中的典型系统互连网络进行了对比，Summit 系统采用的是 Mellanox 的树形互连网络，Piz Daint 系统采用的是 Cray Aries 互连网络. 从表中可以看到，神威 E 级原型机网络链路传输速率最高，达到 28 Gbps；节点上网带宽为 28 GB/s，是神

表 1　神威 E 级原型机和其他系统网络基础性能比较

	TOP500 排名 (2019.11)	链路传输速率/ Gbps	节点上网带宽/ (GB/s)	MPI 带宽/ (GB/s)
神威 E 级原型机	—	28	28	26
Summit (Mellanox)	1	25	25	23
Piz Daint (Aries)	6	14	10.5	9.7
神威·太湖之光	3	14	7	6

威·太湖之光的 4 倍，是 Piz Daint 系统的 2.6 倍，比 Summit 提高 12%；节点 MPI 带宽为 26 GB/s，达到节点上网带宽的 92.8%，高于 Summit 系统的 92% 和 Piz Daint 系统的 92.3%.

5.2 点对点消息性能

由于神威 E 级原型机的 PCI-E 不能完全发挥 SWHNI 芯片的带宽，我们的消息带宽是基于神威 E 级原型机运算节点的升级版本测试的，该运算节点采用 PCIe 接口数据宽度 16 个通道，与 SWHNI 芯片接口匹配.

我们使用不同长度的 RDMAW 消息测试神威 E 级原型机的消息带宽. 测试结果如图 10 所示，1 MB 长度消息的带宽 13.269 GB/s. 事实上，神威互连网络链路采用了 64/66 编码，网络硬件可用的网络端口链路峰值带宽为 13.63 GB/s；消息包最大数据净负荷 97.7%，因此软件可用的网络端口理论峰值带宽为 13.324 GB/s. 由此，1 MB 长度消息实际传输带宽达到了网络端口链路峰值带宽的 97.3%，达到了软件可用的网络端口理论峰值带宽的 99.6%；消息长度为 4 KB 时，消息带宽达到 7.08 GB/s，超过链路峰值带宽的一半.

在消息虚拟化方面，由于采用了虚拟机消息旁路宿主机操作系统及虚拟机操作系统的机制，消息虚拟化的开销非常小，消息延迟相比非虚拟化环境只增加约 0.01 μs，带宽损失低于 0.06%，几乎可以忽略.

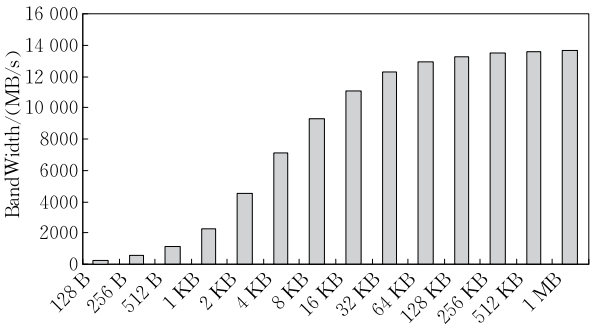

图 10 不同长度消息带宽

5.3 集合消息性能

使用 IMB 标准测试程序，对比测试了部分软件集合操作和基于硬件实现的集合操作的性能. 表 2 和图 11 给出了同步操作软硬件实现的开销，表 3 和图 12 给出了 8 B 小粒度的全归约操作软硬件实现的开销，表 4 和图 13 给出了 1 KB 以下广播操作软硬件实现的开销.

表 2 同步操作(Barrier)软硬件实现延迟对比（单位：μs）

节点规模	实现方式	延迟
256	软件	59.19
	硬件	7.70
512	软件	78.51
	硬件	10.30
1024	软件	98.05
	硬件	13.48

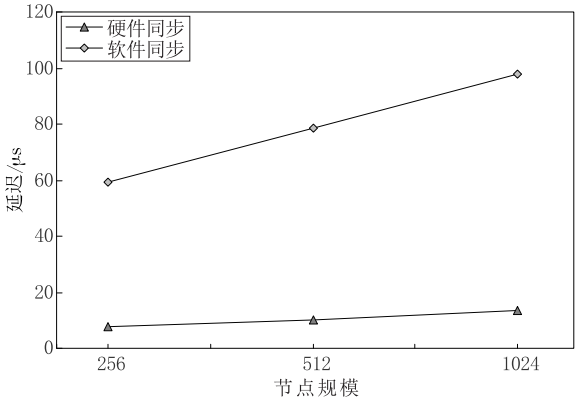

图 11 同步操作软硬件实现开销对比

表 3 全归约操作(Allreduce)软硬件实现延迟对比（单位：μs）

节点规模	实现方式	延迟
256	软件	63.15
	硬件	7.71
512	软件	81.47
	硬件	10.30
1024	软件	108.56
	硬件	13.49

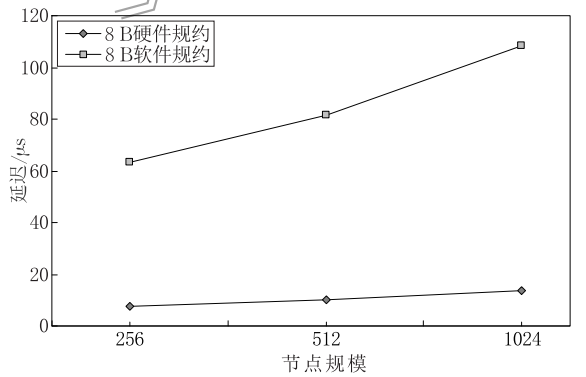

图 12 全归约操作软硬件实现开销对比

表 4 广播操作(Bcast)软硬件实现延迟对比（单位：μs）

节点规模	实现方式	8 B 延迟	128 B 延迟	1024 B 延迟
256	软件	28.43	35.04	38.21
	硬件	7.76	8.05	9.29
512	软件	32.73	41.12	44.40
	硬件	8.37	8.67	9.66
1024	软件	46.65	56.76	61.54
	硬件	11.00	11.23	11.67

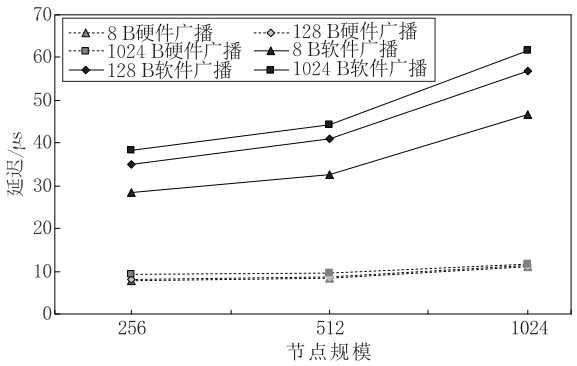

图 13　广播操作软硬件实现开销对比

测试结果表明，在神威 E 级原型机系统上，基于硬件实现的同步和小粒度的广播、全归约性能均优于软件实现，且性能随规模增长变化较小.

5.4　系统计算性能

基于 SW26010+、SWHRC、SWHNI 三款芯片构建的神威 E 级原型机，Linpack 测试性能达到 2.55 PFlops，效率 81.5%，效率在 2018 年中国高性能计算机 TOP100 中位居第一，图 14 给出了 2018 年中国高性能计算 TOP10 的 Linpack 效率对比.

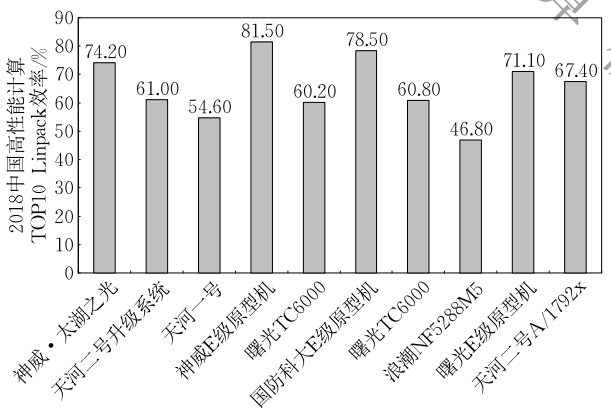

图 14　2018 年中国高性能计算 TOP10 的 Linpack 效率

神威 E 级原型机互连网络为科学计算应用课题提供了高效可靠的消息服务. 表 5 给出了六个科学计算应用在神威 E 级原型系统中运行测试结果，与神威·太湖之光系统相比，六个应用的整机并行效率均有提高.

表 5　典型科学计算的整机并行效率对比

	神威·太湖之光	神威 E 级原型机
全球大气非静力云分辨模拟	50	71
全球高分辨率海浪模式	62	73
纳米材料的导热性质模拟	87	89
中高能粒子输运模型(ADS)	80	95.2
近空间飞行器高超声速绕流	66	75
LSTM 机器翻译大规模训练	90 (1024 节点)	92 (512 节点)

6　结　论

超级计算机即将迈入 E 级计算时代，通过对比国内外各个 E 级计算机研究计划[①]可以看到，E 级计算系统继续延续着提升处理器性能和增加处理器数量相结合的技术路线. 单处理器计算能力超过每秒十万亿次浮点结果，要提供与之平衡的通信能力，其难度和代价远远大于计算能力的提升. 处理器数量超过 10 万个，通信线路达到数百万条，对系统互连网络在可扩展性和可靠性方面提出了新的巨大挑战. 如本文前述，应对挑战，新一代神威互连网络采用了一系列创新技术，在支撑更大互连规模的同时，满足高性能处理器间高带宽、低延时、高可靠的通信需求.

新一代神威互连网络采用 28 Gbps 高速串行传输技术，设计开发了基于 DESDP 架构的双端口高性能网络接口芯片 SWHNI，处理器峰值上网带宽达到神威·太湖之光的 4 倍；提出基于数据流驱动的双轨乱序消息机制和新颖高效的神威消息原语，点到点消息传输性能达到理论峰值的 97.3%；设计实现了可扩展的硬件集合通信，大规模系统集合通信延迟大幅降低；设计开发了 40 端口高阶路由器芯片 SWHRC，实现 20 万处理器互连，网络直径仍然只有 7 步，和 4 万处理器规模的神威·太湖之光相同，可有效支持大规模系统的高效可扩展. 综合运用链路均衡、阻抗控制、高速 PCB 设计、前向纠错、链路重传、路由重构、消息重传等技术，同时在传统树网基础上，提出并实现了一种双轨泛树网络结构，建立了从物理层到传输层的层次化网络容错体系，实现了数据不间断可靠传输，大幅提升了超大规模网络的可靠性. 新一代神威互连网络在应对 E 级计算机通信墙、可靠性墙、可扩展性墙等挑战方面，进行了卓有成效的探索，为 E 级计算机互连网络的研制奠定了基础.

致　谢　非常感谢北京大学杨超教授、自然资源部第一海洋研究所赵伟研究员、中国科学院过程工程研究所侯超峰副研究员、中国科学院近代物理研究所杨磊研究员、中国空气动力研究发展中心李志辉研究员、清华大学杨广文教授在应用通信性能测试方面提供的宝贵帮助，在此再次表示诚挚谢意！

① Lucas R, Ang J, Bergman K, et al. DOE advanced scientific computing advisory subcommittee (ASCAC) report: Top ten exascale research challenges. http://www.osti.gov/scitech/biblio/1222713. 2014

参 考 文 献

[1] Kim J, Dally W J, Abts D. Flattened butterfly: A cost-efficient topology for high-radix networks. ACM SIGARCH Computer Architecture News, 2007, 35(2): 126-137

[2] Derradji S, Palfer-Sollier T, Panziera J P, et al. The BXI interconnect architecture//Proceedings of the IEEE High Performance Interconnects. Santa Clara, USA, 2015: 18-25

[3] Leiserson, Charles E. Fat-trees: Universal networks for hardware-efficient supercomputing. IEEE Transactions on Computers, 1985, C34(10): 892-901

[4] Shanley Tom. InfiniBand network architecture. PC System Architecture, 2003

[5] Liao X K, Pang Z B, Wang K F, et al. High performance interconnect network for Tianhe system. Journal of Computer Science & Technology, 2015, 30(2): 259-272

[6] Chen D, AEisley, Heidelberger P, Senger R M, Satterfield D L, Steinmacher-Burow B, Parker J J. The IBM Gene/Q interconnection network and message unit//Proceedings of the IEEE High Performance Computing, Networking Storage and Analysis. New York, USA, 2011: 26:1-26:10

[7] Faanes G, Bataineh A, Roweth D, et al. Cray Cascade: A scalable HPC system based on a Dragonfly network//Proceedings of the IEEE High Performance Computing, Networking Storage and Analysis. Los Alamitos, USA, 2012: 103:1-103:9

GAO Jian-Gang, M. S., senior engineer. His research interests include computer architecture and high-performance interconnection network.

LU Hong-Sheng, M. S., senior engineer. His research interests include computer architecture and high performance interconnection network.

HE Wang-Quan, Ph. D., senior engineer. His research interests include computer architecture and parallel language design.

REN Xiu-Jiang, Ph. D., engineer. His research interests include computer architecture and high performance interconnection network.

CHEN Shu-Ping, M. S., senior engineer. His research interests include computer architecture and software for interconnection network.

SI Tian-Han, M. S., engineer. His research interests include computer architecture and high performance interconnection network.

ZHOU Zhou, Ph. D., engineer. His research interests include computer architecture and high performance interconnection network.

HU Shu-Kai, M. S., engineer. His research interests include computer architecture and high performance interconnection network.

YU Kang, Ph. D., engineer. His research interest is computer architecture.

WEI Di, M. S., engineer. His research interest is computer architecture.

Background

The high-performance interconnection network and message mechanism are studied on the purpose of independent control.

The communication performance must match the fast developing computing capability on the road to exascale system. The worldwide top supercomputers mainly select Mellanox InfiniBand, Cray Aries, Intel Onmi-path, and employ the 25 Gbps transmission technique to implement their interconnection network. The networks of the top domestic supercomputer, such as "Sunway Taihu Light" and "Tianhe 2", are constructed based on 14 Gbps transmission. Our team makes the break-through on the key technologies of 28 Gbps transmission, high-radix router, high-performance network interface, high-efficient and reliable network architecture. Furthermore, Sunway network chipset of new generation is designed, and the network of Sunway exascale prototype system is constructed. They all contribute to the design of the domestic exascale supercomputer.

This project is supported by the National Key R&D Project "Verification System of the Key Techniques for the Exascale System" under Grant No. 2016YFB0200500. The research achieves the goal of innovative design of the exascale system by constructing the large-scale verification system, mastering the techniques of new interconnection network architecture, and testing based on domestic components and parts. The research team has long-term technology accumulation in high-performance computing and accomplished "Sunway Taihu Light" supercomputer in 2015, supported by the 863 project. It won the first place of Top500 for 4 consecutives times.