doi:10.3772/j.issn.1002-0470.2014.02.001

基于紧耦合单跳步多平面架构的高端服务器设计①

王恩东② 陈继承③ 胡雷钧 公维峰

(浪潮集团有限公司 北京 100085)
(高效能服务器和存储技术国家重点实验室 北京 100085)

摘 要 针对高端服务器设计面临的可扩展性问题,提出了一种紧耦合单跳步多平面(TSMP)体系结构设计方法。该方法采用双侧多平面互连结构,支持8 –32 路规模无缝扩展;基于两级目录结构的高速缓存一致性实现方法,支持高并发一致性访问和高效冲突处理,有效降低一致性访问传输、处理延迟。该方法已应用于浪潮32 路K1 高端服务器的设计,对设计的系统进行了内存性能、处理性能和可扩展性测试,测试结果表明,采用该设计可使高端服务器的计算、访存性能随系统规模从单路到32 路线性增长。K1 高端服务器支持基于QPI1.0 协议的Intel 安腾(Itanium)4 核CPU-Tukwila 和8 核CPU-Polson,是中国研制的首台投入商业化应用的高端服务器。

关键词 缓存一致性非均匀存储访问(CC-NUMA),紧耦合单跳步多平面(TSMP),QPI 协议,cache 一致性,目录cache

0 引 言

保持系统性能随规模线性增长是高端服务器设计面临的关键挑战。对于传统前端总线(front side BUS,FSB)结构处理器组成的高端服务器系统[1-4],由处理器协同芯片(或北桥芯片)集中管理内存并维护全局高速缓存(cache)一致性。由于FSB 带宽限制,高端服务器系统性能受到严重制约。随着计算机技术的发展,点到点互连方式和内置存控技术引入处理器设计,每个处理器都集成内存控制器并外接内存[5],都在全系统空间上管理一段一致性内存空间。相比FSB 结构集中式管理内存方式,此时处理器协同芯片功能聚焦在跨节点远端内存访问时的全局cache 一致性维护,而节点内跨处理器本地内存访问不需要通过处理器协同芯片处理,从而可有效提升局部访存性能。因此,点到点的互连方式和内置存控技术带来了高端服务器设计的重大变革[6,7]。

必须看到,基于点到点互连方式组成的高端服务器系统存在访存不均衡问题,内存管理分布化使跨节点内存访问比FSB 结构可能需要更多级跳步(hop),访存延迟相对增加,尤其在对远端数据的访问需要多次缓存一致性操作才能完成的情况下,跨节点访问的效率进一步降低,从而使系统难以保持其性能与规模的线性关系,因此高端服务器系统访存的非均匀性使得系统的可扩展性和内存墙问题显得突出。访存不均衡还会导致负载不均衡,从而进一步降低系统性能。故对于点到点互连高端服务器系统,由于访存和负载的非均匀性,如何解决系统的可扩展性从而保障其计算能力和访存能力随规模线性增长,则成为高端服务器设计面临的关键问题。本文围绕32 路高端服务器设计,提出了紧耦合单跳步多平面(tightly-coupled single-hop multi-plane,TSMP)高端服务器体系结构设计方法,满足了高端服务器设计的高可扩展性和高并发处理能力需求,保障了系统性能随规模线性增长。

1 32 路高端服务器TSMP 体系结构设计

本研究在32 路高端服务器TSMP 体系结构设计上,注重了两个方面:首先在结构拓扑设计上,提

① 863 计划(2008AA01A202)和973 计划(2010CB735905)资助项目。
② 男,1966 年生,研究员;研究方向:计算机体系结构,信息存储系统体系结构,系统总线协议技术等;E-mail:wangend@inspur.com
③ 通讯作者,E-mail:chenjch@inspur.com
(收稿日期:2013-04-01)

出了双侧多平面互连结构，兼顾高可靠性和高扩展性需求，实现互连网络间单跳步（single hop），可以无缝支持 8 – 32 路结构扩展，同时由于每个平面网络都能连接所有节点，故只要有一套平面网络正常工作就可保证系统连通；然后在全局缓存一致性设计上，提出了基于两级目录结构的 cache 一致性设计思想及硬件实现方法，支持高并发一致性访问和高效冲突处理，有效降低跨节点访问传输延迟和阻塞延迟。以下进行详细论述。

1.1　双侧多平面互连结构

基于 TSMP 结构的 32 路高端服务器结构拓扑如图 1 所示，其中“S”为 CPU Socket，可为 Intel Itanium Tukwila 4 核处理器或 Polson 8 核处理器，“CC”为支持快速通道互连（QPI）接口的处理器协同芯片，“HR”为高速互连芯片。

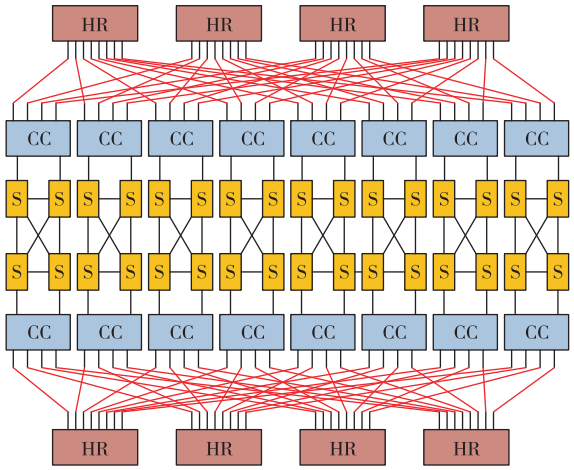

图 1　TSMP 架构的 32 路高端服务器结构拓扑

1.2　TSMP 结构拓扑建模分析

1.2.1　TSMP 互连结构与 HR 芯片的关系

节点间互连网络性能对高端服务器系统有重要影响，为定量评估 TSMP 架构与高端服务器系统性能之间的关系，本文构建了 TSMP 互连网络模型，采用时钟周期驱动模式模拟网络中包传送过程及各节点的处理和发送过程，完成了传输瓶颈分析和不同拓扑结构（单 HR、双 HR 和 4HR）的性能对比，包括以下几个方面：

（1）激励来源

输入激励来自于处理器最后一级高速缓存（last level cache，LLC）的 miss，如 Load/Store 指令执行时在 LLC 的 miss、Store 指令在 L1-LLC 存储层次上无独占权限的 Hit 及 Cache 替换、写回产生的交互消息；通过 IA64 指令级多核模拟器，生成访问 Trace，并基于该 Trace 特征，人工生成激励。

（2）模拟流程

将 TSMP 互连结构抽象成有向简单图，使用邻接表配置连接关系，并指定每个图上节点类型，目前节点类型有 L3（LLC）、CC 和 HR。具体实现时，使用 Node 抽象类，共用消息机制；在具体模拟策略上，网络各节点行为参数可调，如 CC 设计简化，以 FIFO 形式双工地对包进行非一致性转发并加以延时，另外 HR 虚信道数目可配，负载强度可调。

（3）评估标准

本文测试以报文平均传输时延和平均阻塞时延来定量分析不同拓扑结构（单 HR、双 HR 和 4HR）的差异关系。

（4）评估结果

TSMP 互连网络模型评估结果如图 2 和图 3 所示（其中横坐标“负载强度”表示报文收发间隔，1/20000表示 20000 时钟周期收发一次报文）。

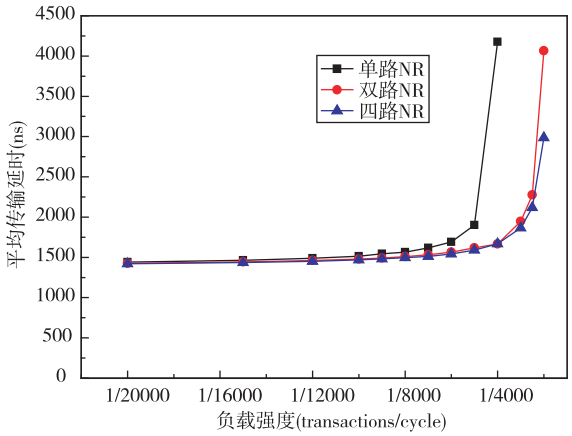

图 2　不同 HR 数目下系统平均传输时延

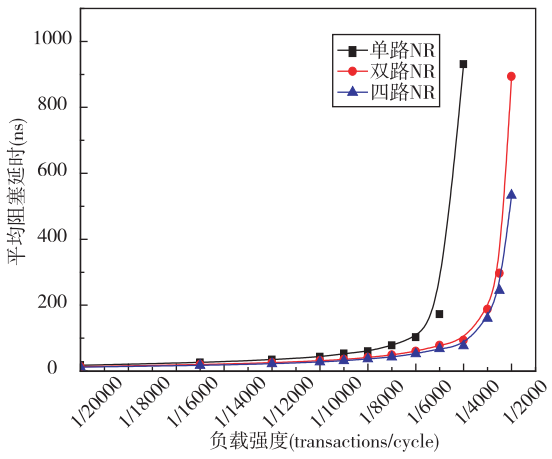

图 3　不同 HR 数目下系统平均阻塞时延

从图 2 和图 3 的仿真评估结果可知,在负载比较轻的情况下,多路 HR 互连芯片在平均传输和阻塞延迟上的优势不明显,但随着负载增加,尤其当网络接近满载时,HR 数目对系统性能影响急剧上升;与单 HR 相比,4HR 显著改善了高负载情况(负载为 1/4000)下平均数据传输时延,有约 150% 改善;与 2HR 相比亦有约为 40%(负载为 1/2000)改善。同样,4HR 下平均阻塞时延与单 HR、2HR 相比更分别有约 500% 和 70% 的改善;因此本文 TSMP 架构采用 4HR 多平面互连结构可以有效降低系统报文平均传输和阻塞时延。

1.2.2 TSMP 互连结构传输报文格式与平均阻塞/传输时延比率的关系

对于本文研究的有阻塞互连网络系统,给出了基于 TSMP 架构的互连网络模型的仿真结果,图 4 所示为不同报文长度情况下系统传输网络的平均阻塞/传输时延比率关系。

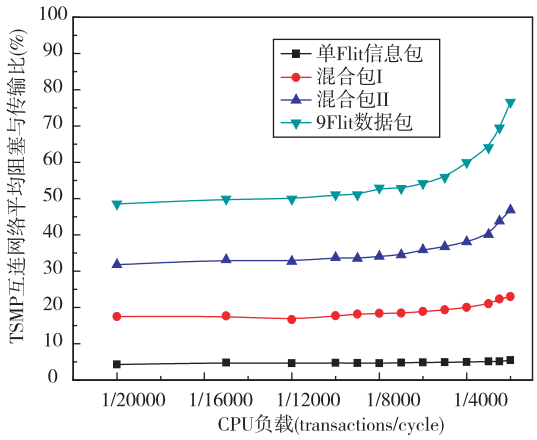

图 4 TSMP 互连网络平均阻塞/传输时延比率

图 4 中横坐标为负载情况,纵坐标表示在不同报文长度情况下的系统平均阻塞时延与平均传输时延的比率。假定报文最小长度单位为 Flit(80bit),其中信息包长为 1Flit,数据包长为 9Flit,则可分为如下 4 种情况:

(1)单 Flit 信息包情况

最下方拟合线为全部为单 Flit 信息包情况下平均阻塞/传输时延比率,从轻负载到重负载其比率一直维持在 5% 左右。

(2)混合包 I 情况

中间偏下拟合线为混合包 I(80% 单 Flit 信息包,20% 的 9Flit 数据包)情况下平均阻塞/传输时延比率,在轻、中度负载情况下,比率约为 17% ~18%,在重负载情况下比率约为 21% ~23%。

(3)混合包 II 情况

中间偏上拟合线为混合包 II(50% 单 Flit 信息包,50% 的 9Flit 数据包)情况下平均阻塞/传输时延比率,在轻、中度负载情况下,比率约为 32% ~35%,在重负载情况下比率约为 40% ~47%。

(4)9Flit 数据包情况

最上方拟合线为全 9Flit 数据包情况下的平均阻塞/传输时延比率,在轻、中度负载情况下,比率约为 50% ~55%,从中度负载到重度负载则呈线性上升,从 60% 快速上升到 80%。

由此可得,节点间互连网络平均阻塞时延与报文长度有直接关系,采用短报文传输则可有效降低其平均阻塞时延,因此节点间互连网络应尽量采用短报文传输并提升其有效载荷。

基于 TSMP 互连网络模型分析,本文制定了 TSMP 架构节点间域 NI 报文组包方式。

1.3 全局缓存一致性设计

1.3.1 两级域 Cache 一致性协议和两级目录设计

Cache 一致性协议层级及目录结构是影响高速缓存一致性非均匀存储访问(cache coherence non-uniform memory access,CC-NUMA)架构计算机系统扩展性和性能的重要因素。中小规模系统可采用单级 cache 一致性协议和单级目录结构,但大规模系统目录层级设计需要在系统扩展性、硬件实现代价和执行效率上取得权衡[8]。

本文 32 路高端服务器采用两级 cache 一致性协议,第一级为节点内 cache 一致性协议,采用 Intel QPI 协议;第二级为节点间 cache 一致性协议,采用自主(NI)协议。

对于 NI 协议,在信道设计上采用 8 虚信道设计,合理分配高速缓存一致性(cache coherence,CC)与非 CC 事务、大数据与小数据事务,保证协议报文的传输高效和无死锁;在数据转发上采用支持干净独占和脏数据转发,有效减少协议报文数量和加速读事务完成;在重试机制上支持协议报文 Nack 和 Retry,保证冲突消解有序进行,避免协议出现死锁和活锁;在错误目录恢复机制上协议实现上通过静默、全监听和解静流程,支持目录在出错情形下自动恢复,提高协议自身容错能力。

同时,本文 32 路高端服务器采用两级协议目录设计方法,其中第一级为根节点(home node)协议目录,第二级为根处理器(home processor)协议目录。对于根节点协议目录,采用全地址目录方案,从而消

除目录替换、目录项写回开销，提升系统的协议处理效率，而根处理器协议目录采用有限目录设计方法降低处理器内部协议目录存储开销，从而实现系统扩展性、硬件开销和性能的均衡。

1.3.2　高并发一致性设计

高端服务器系统中 cache 一致性请求项高并发访问和冲突处理机制尤为重要，本文的高并发一致性设计策略包括：

（1）多协议流水线机制

处理器协同芯片内部实现了 10 条协议流水线，用于处理 CC 与非 CC 事务，各条流水线之间完全独立，有效提高协议报文的处理速度。

（2）采用目录 cache 机制

将最近最常使用的目录项缓存在目录 cache 中，利用数据访问的局部性降低目录信息的访问延迟，提高协议的处理性能；目录 cache 分为 4 个 Bank，每个 Bank 实现 8 路组相联，最大支持 32 个目录失效请求。

（3）失效引擎机制

采用失效引擎机制加速失效集合提取和失效报文生成，并负责统一集中发送失效报文，缩短失效报文的产生时间，同时提高失效报文的发送效率。

（4）分区控制机制

采用分区（partition）控制机制减少不必要的失效报文，一方面提高协议的处理性能，另一方面也减少对通信带宽的浪费。

（5）冲突处理解决机制

采用请求冲突检测机制检测出系统运行过程中出现的所有请求冲突情况，并实时采用请求冲突化解机制将冲突请求的处理串行化，从而解决一致性请求冲突的情况。

1.4　处理器协同芯片设计

处理器协同芯片时钟域和规格如图 5 和图 6 所示，主要有 4 个时钟域，即 QPI 接口时钟域、内存接口时钟域、内核时钟域及网络接口时钟域。通过优化逻辑设计，实现了节点连接器芯片内核频率 400MHz。

处理器协同芯片结构复杂，逻辑庞大，超过上亿晶体管，对芯片后端设计带来较大挑战。本文通过预布局及多次布局优化方法最终以 18mm × 18mm 的面积规模实现了 90nm 工艺处理器协同芯片设计，一次流片成功。

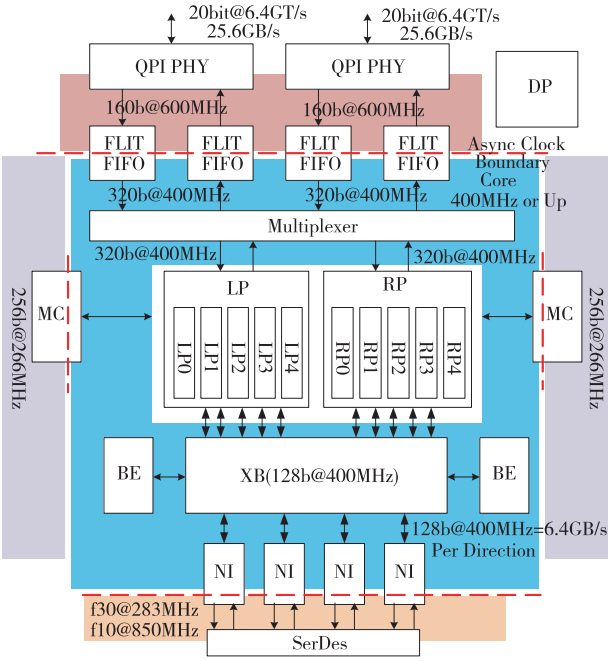

图 5　处理器协同芯片时钟域划分

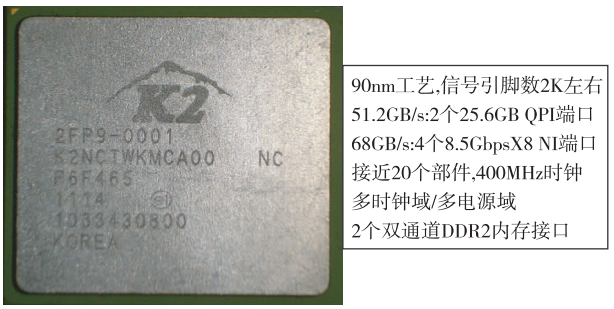

图 6　处理器协同芯片规格参数

2　性能测试

为验证基于 TSMP 架构的 32 路高端服务器的计算、访存性能与规模的线性关系，对其开展了内存性能测试、处理性能测试和可扩展性测试等。

2.1　内存性能测试

内存性能测试采用 Stream 测试工具，测试结果如图 7 所示。由图 7 可见，32 路高端服务器系统 Stream 内存性能从单 CPU（1 CPU/4 核）到 32 路（32 CPU/128 核）接近线性增长。

2.2　处理性能测试

采用 SPEC CPU 2006 测试多处理器处理性能（rate 模式），测试结果如图 8 所示。

从图 8 可见，32 路高端服务器 SPEC CPU 定点计算性能从单 CPU（1 CPU/4 核）到 32 路（32 CPU/128 核）接近线性增长。

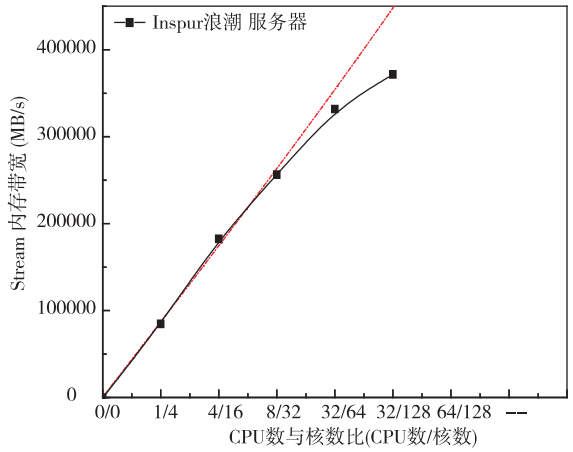

图 7　Stream 内存性能测试结果

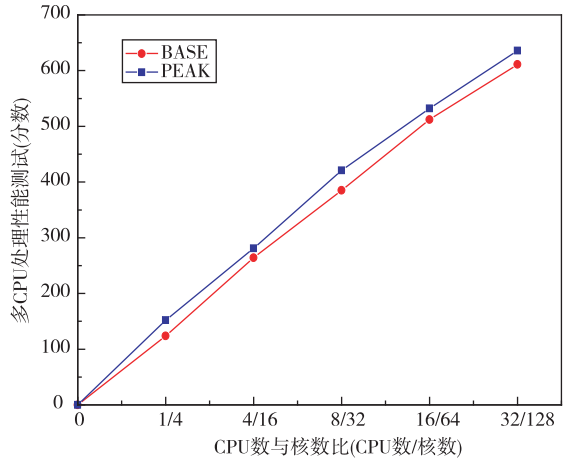

图 8　多 CPU 处理性能测试结果

2.3　可扩展性测试

考察系统整机进行科学应用计算的处理能力及随 CPU 核数增加的系统扩展能力,测试结果如图 9 所示。

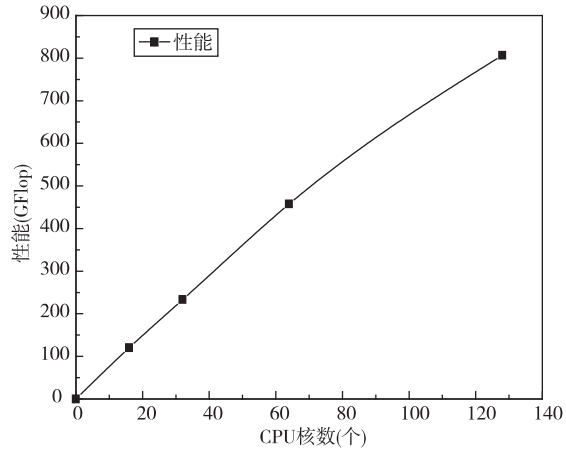

图 9　可扩展性测试结果

从图 9 可见,32 路高端服务器浮点计算能力达到 806.6GFlops,从单 CPU(1 CPU/4 核)到 32 路(32 CPU/128 核)其性能接近线性增长。

32 路高端服务器作为处理主机已在建设银行新疆分行区域平台特色业务处理中实现了示范应用。应用结果表明,基于 TMSP 架构的 32 路高端服务器能够可靠承载银行类中间业务平台应用。

3　结　论

本文针对处理器采用点到点互连方式和内置存控技术对高端服务器系统设计带来的挑战,提出了一种基于紧耦合单跳步多平面(TSMP)架构的高端服务器设计方法,该方法通过双侧多平面互连拓扑结构支撑 8 – 32 路规模无缝扩展,采用基于两级目录结构的全局 cache 一致性硬件维护方法有效支持高并发一致性访问和高效冲突处理。内存性能、处理性能和可扩展性测试结果表明,采用本文方法设计的浪潮 32 路 K1 高端服务器的系统计算和访存能力可随系统规模从单路到 32 路线性增长。

参考文献

[1] Laudon J, Lenoski D. The SGI Origin: a ccNUMA highly scalable server. In: Proceedings of the ACM 24th Annual International Symposium on Computer Architecture, New York, USA, 1997. 241-251

[2] Gostin G, Collard J-F, Collins K. The architecture of HP superdome shared-memory multiprocessor. In: Proceedings of the ACM 19th Annual International Conference on Supercomputing, New York, USA, 2005. 239-245

[3] Aono F, Kimura M. The AZUSA 16-way Itanium server. *IEEE Micro*, 2000, 20(5): 54-60

[4] Gharachorlooy K, Sharma M, Steely S, et al. Architecture and design of AlphaServer GS320. *ACM Sigplan Notices*, 2000, 35(11): 13-24

[5] Conway P, Hughes B. The AMD Opteron northbridge architecture. *IEEE Micro*, 2007, 27(2): 10-21

[6] Feehrer J, Rotker P, Shih M, et al. Coherency hub design for multisocket SUN servers with coolthreads technology. *IEEE Micro*, 2009, 29(4): 36-47

[7] Kota R, Oehler R. HORUS: Large-scale symmetric multiprocessing for Opteron system. *IEEE Micro*, 2005, 25(2): 30-40

[8] Acacio M. E, Gonzalez J, Garcia J. M, et al. A two-level directory architecture for highly scalable cc-NUMA multiprocessors. *IEEE Transactions on Parallel and Distributed Systems*, 2005, 16(1): 67-79

Design of high-end server based on tightly-coupled single-hop multi-plane architecture

Wang Endong, Chen Jicheng, Hu Leijun, Gong Weifeng
(Inspur Group Co., Ltd., Beijing 100085)
(State Key Laboratory of High-end Server & Storage Technology, Beijing 100085)

Abstract

A new method based on the tightly-coupled single-hop multi-plane (TSMP) architecture is put forward to solve the system expansibility issue in high-end server design. It adopts a two-side single-hop multi-plane topology to support the seamless extending of system from 8 to 32-way. A two-tier directory based cache coherence maintenance method is applied to sustain both high parallel cache coherence requests and high efficiency conflicts handling, which reduces the system transmission delay and disposal delay remarkably. The TSMP architecture has beea used in the design of an Inspur K1 high-end server. Testing results of Stream, SPEC CPU and high pertormance Linpack (HPL) indicate that K1 system performance increases linearly from 1 to 32-way. The K1 high-end server supports the QPI1.0 based Intel Itanium 4-core Tukwila and 8-core Polson CPUs. It is China's first independent commercial high-end server.

Key words: cache coherence non-uniform memory access (CC-NUMA), tightly-coupled single-hop multi-plane (TSMP), QPI, cache coherence, cache directory