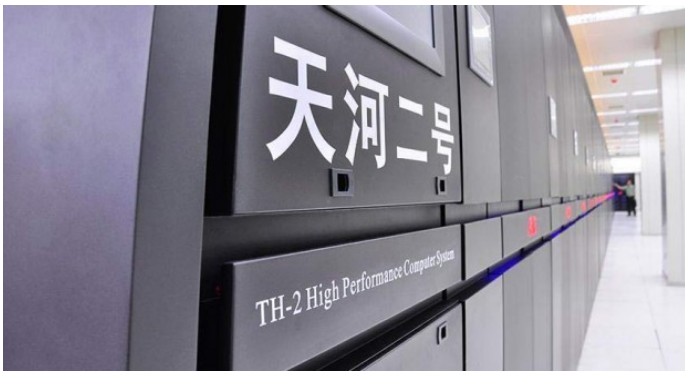
天河二号

随着 2013 年 6 月 17 号 Top500 排行榜的公布，国防科技大学研制的天河二号成为全球运算速度最快的计算机。这是继 2010 年 11 月份天河一号排名世界第一之后，中国的超级计算机再登世界超算之巅。本文将以 Jack Dongrra 教授的报告为基本材料，在介绍天河二号的大体信息同时，着重介绍技术上的细节以及相关的性能指标。



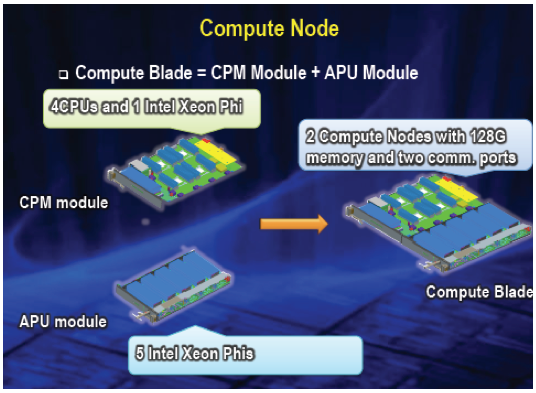
天河二号，英文名称为 Tianhe-2 或者 Milky Way-2，是由国防科技大学等单位研制的异构超级计算机。将于 2013 年底落户位于广东省的国际超级计算广州中心。届时，天河二号将对外开放接受运算任务，用于实验、科研以及教育领域。

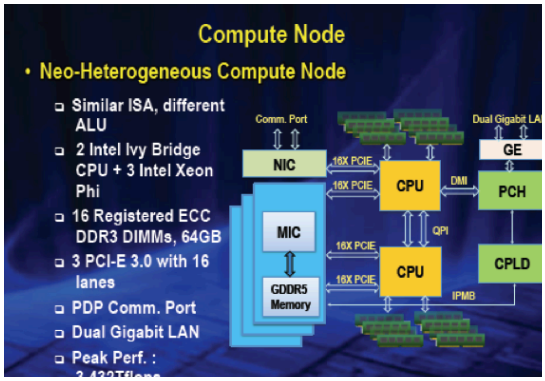
天河二号的理论峰值性能将达到 54.9petaFLOPS。主要由 Intel 的 Ivy Bridge Xeon E5 处理器以及 Xeon Phi 协处理器通过中国自主研发的 TH-Express-2 高速网络互联起来。其中 E5 处理器为 32, 000 个，Xeon Phi 协处理器为 48, 000 个，一共由 3,120,000 个核组成。据说，在计算性能上能够超过天河二号的超级计算机，要等到 2015 年才有可能部署。

计算节点

每个计算节点由 2 块 Xeon E5-2692 处理器以及 3 块 Xeon Phi 协处理器组成。每两个结点组成一个 board，每 16 个 board 组成一个 frame，4 个 frame 组成一个 rack。而整个系统由 125 个 rack 组成（如下图所示）。

其中，每个 board（也既图中的 blade）被划分成了 CPM 模块和 APU 模块。CPM 模块是 4 个 CPU 加上 1 个 Xeon Phi 组成。APU 模块是 5 块 Xeon Phi 组成。E5 处理器同 Xeon Phi 之间采用的是 PCI-E 3.0 的总线，不过由于 Xeon Phi 只支持 PCI-E 2.0，限制了其数据交换能力。Board 的结构如下图所示：





天河二号使用的 Xeon Phi 是有 57 个核的众核协处理器，时钟频率为 1.1GHz。通常，Xeon Phi 有 61 个核。而由于天河二号使用的是早期产品，存在流片的一些问题。如果启用全部核心，会存在运算周期协调之间冲突问题，所以屏蔽掉了 4 个核。对于每个计算结点，2 个 E5 处理器提供了 2*0.2112teraFLOPS 的浮点数运算性能。而 3 块 Xeon Phi 提供了 3*1.003 teraFLOPS 的服点运算性能，总计高达 3.431 teraFLOPS 每节点。整个系统有 16,000 个节点，因此可以提供约为 54.9 petaFLOPS 的处理性能。

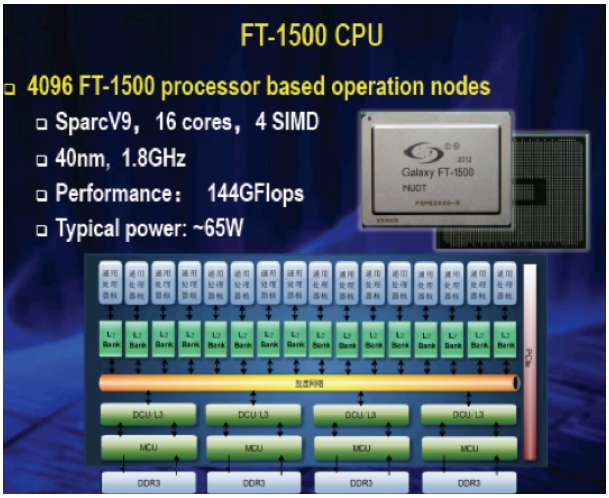
每个节点上的内存大小为 64GB，而每个 Xeon Phi 协处理器上拥有 8GB 的内存。因此每个计算结点的内存大小为 88GB。16,000 个结点使得天河二号的内存总量达到 1.404PB。

功耗和冷却系统

在带载情况下，系统的功耗为 17.6MW。所使用的冷却系统为水冷系统。如果算上冷却系统，总功耗将高达 24MW。在天河二号的机柜表面，有条状的灯光。其颜色将根据机器正在运行时的实际功耗而变化颜色。是不是很酷啊。

前端处理器

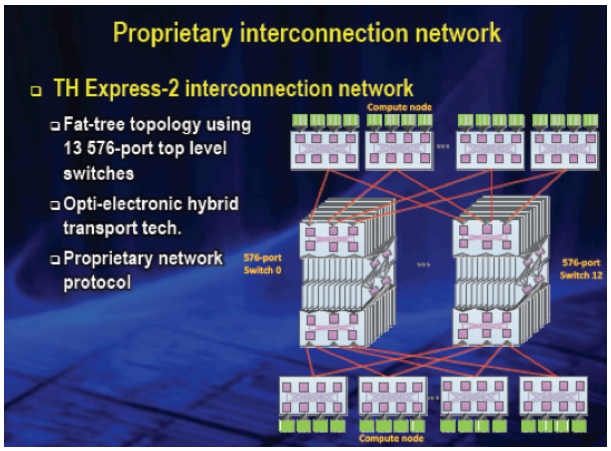
除了计算结点之外，天河二号采用了 4096 个 Galaxy FT-1500 处理器。这些处理器是国防科大自主设计研发的，但并不是天河二号的一部分。FT-1500 是基于

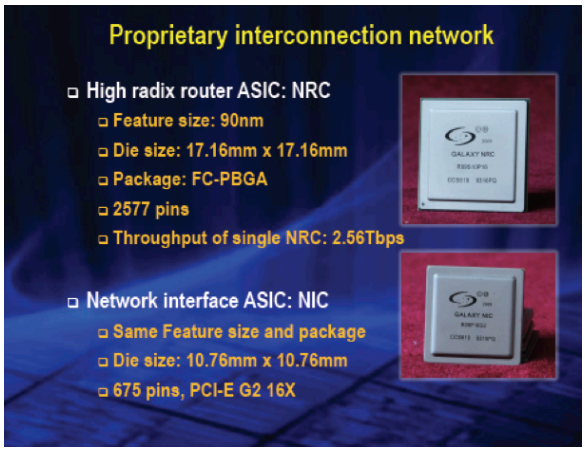


SparcV9 的 16 核处理器。使用了 40nm 的技术，运行频率为 2.2GHz。峰值性能为 211gigaFLOPS。前端处理器的主要用作任务调度。

互联

天河二号采用的国防科大自己的专有互联技术，使用光电混合传输技术（Optoelectronics Hybrid Transport Technology），称之为 TH Express-2。整个系统以 13 个大型交换机，每个拥有 576 个端口。其中使用了自主研发的控制芯片 NRC。该芯片使用 90nm 的工艺，单个控制器的吞吐量高达 2.56 Tbps。在一项通过 MPI 在 12,000 个节点间，广播 1K 的数据的实验中，耗时仅为 9 us。

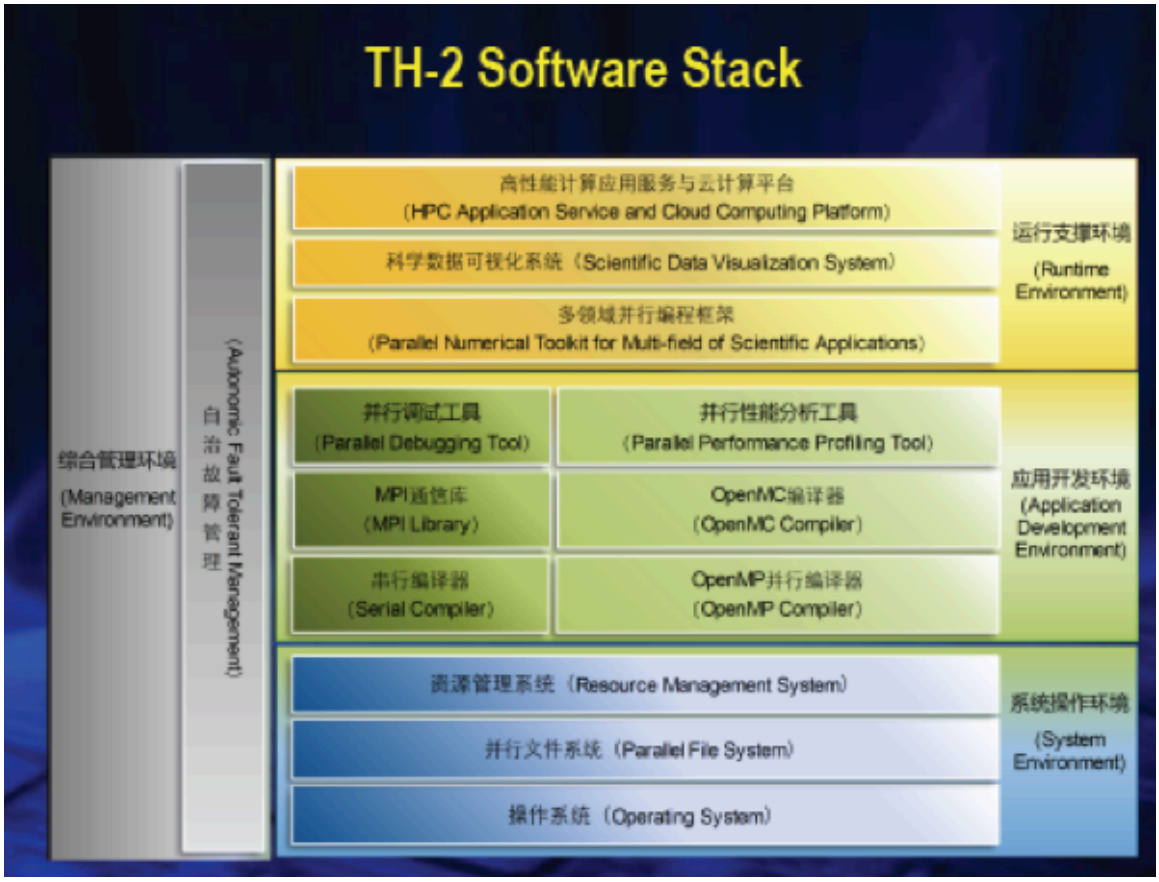




软件栈

天河二号使用的是麒麟（Kylin）Linux 操作系统，与其他主流操作系统相兼容。其上运行的是 SLURM 作为资源管理与任务调度。所使用的编译器包括 Fortran，C，C++，Java，OpenMP 以及根据 MPI 3.0 协议实现的 MPICH。有意思的是，国防科大正在开发一个名为 OpenMC 的编程模型。该编程模型能够为 CPU 以及 Xeon Phi 协处理器的软硬件提供统一的逻辑层。该编程模型正在开发中。天河二号使用了 Intel 的 ICC 编译器，同时自行开发了基于 Intel MKL 以及 BLAS 的高效数学函数库。

天河二号的存储容量高达 12.4PB，使用了 H2FS 文件系统。该文件系统是国防科大自行开发的，相关论文已经发表在 ISC 13 上。



最后我们用两张表来总结天河二号的各项性能指标：

Summary of the Tianhe-2 (TH-2) or Milkyway-2	
Items	Configuration
Processors	32,000 Intel Xeon CPU's + 48,000 Xeon Phi's (+ 4096 FT-1500 CPU's frontend) Peak Performance 54.9 PFlop/s (just Intel parts)
Interconnect	Proprietary high-speed interconnection network, TH Express-2
Memory	1 PB
Storage	Global Shared parallel storage system, 12.4 PB
Cabinets	125 + 13 + 24 = 162 compute/communication/storage cabinets
Power	17.8 MW
Cooling	Closed air cooling system

Summary of the Tianhe-2 (TH-2) Milkyway-2

Model	TH-IVB-FEP
Nodes	16000
Vendor	NUDT, Inspur
Processor	Intel Xeon IvyBridge E5-2692
Speed	2.200 GHz
Sockets per Node:	2
Cores per Socket:	12
Accelerator/CP:	Intel Xeon Phi 31S1P
Accelerators/CP per Node:	3
Cores per Accelerator/CP:	57
Operating System:	Kylin Linux
Primary Interconnect:	Proprietary high-speed interconnecting network (TH Express-2)
Peak Power (MW):	17.8
Size of Power Measurement (Cores)	3,120,000
Memory per Node (GB)	64

Summary of all components

CPU Cores	384,000
Accelerators/CP	48,000
Accelerator/CP Cores	
Memory	1,024,000 GB