日本Oakforest-PACS超级计算机概览

● 陈皖苏　邱家权
江南计算技术研究所 无锡 214083

摘要：
2016年12月，日本Oakforest-PACS超级计算机首次进行全系统运转，其峰值性能为25Pflops，从而超越“K（京）”系统，正式成为目前日本运算速度最快的超级计算机。本文主要对Oakforest-PACS系统的软、硬件配置及其技术特点等加以简要论述。

关键词：Oakforest-PACS，超级计算机，互连网络，冷却系统

1. 引言

2016年11月中旬，第48届全球超级计算机TOP500排行榜发布，日本先进高性能计算联合中心（JCAHPC）运营的超级计算机Oakforest-PACS位居第六。第48届TOP500排行榜的前10台系统情况如表1所示。

表1 第48届TOP500排行榜前10台系统

排名	*系统名称 制造商/时间*	*安装地点*	*峰值性能 实测性能 (万亿次)*	*功耗 (千瓦)*	*能效 (Mflops/W)*	*效率 (%)*
1	神威·太湖之光（Sunway TaihuLight） 中国国家并行计算机工程技术研究中心/2016	中国国家超级计算无锡中心	125435.9 93014.6	15371	6051.3	74.15
2	天河二号（Tianhe-2） 中国国防科大/2013	中国国家超级计算广州中心	54902.4 33862.7	17808	1901.54	61.68
3	泰坦（Titan） Cray公司/2012	美国橡树岭国家实验室	27112.5 17590	8209	2142.77	64.88
4	红杉（Sequoia） IBM公司/2012	美国劳伦斯利弗莫尔国家实验室	20132.7 17173.2	7890	2176.58	85.3
5	科里（Cori） Cray公司/2016	美国劳伦斯利弗莫尔国家实验室	27880.7 14014.7	3939	3557.93	50.26
6	Oakforest-PACS 富士通/2016	日本先进高性能计算联合中心	24913.5 13554.6	2719	4985.69	54.4
7	京（K computer） Fujitsu公司/2011	日本理化研究所	11280.4 10510	12660	830.18	93.17
8	代恩特峰 （Piz Daint） Cray公司/2016	瑞士国家超级计算中心	15988 9779	1312	7453.51	61.2
9	米拉（Mira） IBM公司/2012	美国阿贡国家 实验室	10066.3 8586.6	3945	2176.58	85.3
10	三位一体（Trinity） Cray公司/2015	美国洛斯阿拉莫斯国家实验室/圣地亚国家实验室	11078.9 8100.9	4233	1913.92	73.12

Oakforest-PACS是一台富士通PRIMERGY CX1640 M1机群，峰值运算性能为每秒25Pflops，其LINPACK测试性能则为13.6Pflops，采用了具有68个核心的Intel“骑士登陆”Xeon Phi 7250处理器。2016年12月初，Oakforest-PACS首次进行全系统运转，成为目前日本运算速度最快的超级计算机系统。Oakforest-PACS系统的外观如图1所示。

2. Oakforest-PACS系统概况

如表2所示，由先进高性能计算联合中心（JCAHPC）研制、运营并由富士通公司构建并的Oakforest-PACS大规模并行集群式超级计算机由8,208个计算节点组成，使用基于英特尔公司的Intel Xeon Phi高性能处理器和基于多核处理器技术的Knights Landing架构，节点间通过Omni-Path连接，LINPACK性能约为K系统的2.2倍。该系统安装于东京大学柏木校区的信息技术中心，由东京大学和筑波大学共同进行运营维护。

图1 Oakforest-PACS系统外观

表3 Oakforest-PACS系统基本构成

性能	*峰值性能*		25Pflops
	实测性能		13.6Pflops
系统效率			54.4%
节点数			8208
总机柜数			102
计算节点	基本构成		8块CX1640 M1集成在PRIMERGY CX600 M1（2U）中
	处理器		Intel Xeon Phi 7250、68核、1.4GHz
	存储器	高带宽	16GB MCDRAM，实际性能490GB/s
		低带宽	96GBDDR4-2400，峰值性能115.2GB/s
互连网络	基本构成		Intel Omni-Path Architecture
	链接速度		100Gbps
	拓扑结构		全对分带宽胖树结构
总功耗			4.2MW（含冷却）

Oakforest-PACS系统的研发特点主要体现在以下三个方面：

（1）日本首次尝试不同超级计算机系统的技术融合

Oakforest-PACS是日本JCAHPC研制的超级计算机。JCAHPC于2013年成立，由筑波大学计算科学研究中心和东京大学信息基础中心组成，直接接受日本政府资助。筑波大学计算科学研究中心的PACS和东京大学信息基础中心的Oakforest-fx都是日本知名的超级计算机系统。Oakforest-PACS正是在这两个系统的基础上研发成功的，研发过程中，两大研发中心打破了以往各自为政的封闭局面，进行了技术、经验、人员及资源的全方位共享，为日本各超级计算机系统的技术融合做出了尝试。

（2）系统搭建着眼实用性

Oakforest-PACS是一个采用开放性尖端技术构建的超并行计算机集群。为了便于更多的使用者使用，该系统并没有使用最先进的处理器，也没有使用图形处理器（GPU），而是多采用已成熟的技术，不过分追求高峰值性能，更注重对以往技术的继承与实用价值的体现。

（3）有利于发挥规模优势

由于拥有众多的计算单元，Oakforest-PACS系统能够完成一些超大规模的单一工作，并可应用于众多领域，目前日本能够达到这一应用水平的超级计算机并不多见。

3. Oakforest-PACS系统硬件

（1） 计算节点

Oakforest-PACS系统采用Intel公司的Xeon Phi 7250处理器（开发代码：Knights Landing），68核、1.4GHz。每块富士通公司研发的CX1640 M1型插件板上配置一块处理器板。每8块插件板整合在富士通公司开发的PRIMERGY CX600 M1（2U）机架中，构成一个计算节点抽屉，如图2所示。该计算节点的内存分为两种，高带宽采用MCDRAM，容量为16GB，实

际性能可达到490GB/s；低带宽采用DDR4-2400，容量为96GB，峰值性能为115.2GB/s。

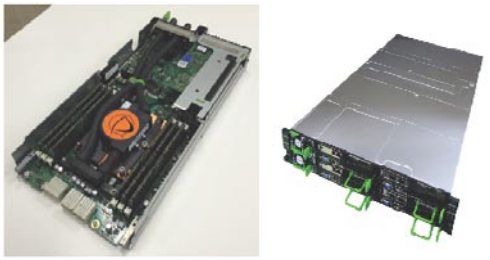

图2 Oakforest-PACS系统的计算节点构成

（2）互连网络

如图3所示，Oakforest-PACS系统的互连网络采用了Intel公司的Omni-Path Architecture，链接速度100Gbps。拓扑结构为全对分带宽胖树网络（Full Bisection Band Fat-tree Net）。该网络拥有768端口的导向器交换机（768 port Director Switch）12台，48端口的边缘交换机（48 port Edge Switch）362台。使用该网络虽费用较高，但可在全系统应用时实现较高的并行性能；当应对不同任务时，计算单元在分割使用方面也能实现更高的自由度。

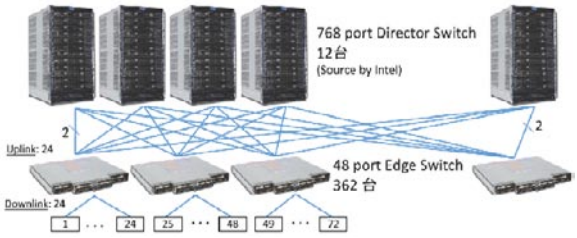

图3 采用了Omni-Path Architecture的全对分带宽胖树网络

（3）并行文件系统

Oakforest-PACS的并行文件系统采用了Data Direct Networks SFA14KE，型号为Lustre File System，总容量26.2PB，总带宽500GB/s。

（4）高速文件缓存系统

Oakforest-PACS的高速文件缓存系统采用Data Direct Networks IME14K，型号为Burst Buffer，Infinite Memory Engime（by DDN），总容量为940TB（NVMe SSD，含电池），总带宽高达1560GB/s。

4. Oakforest-PACS系统软件

Oakforest-PACS系统的软件配置十分丰富，如表3所示。

表3 Oakforest-PACS系统的软件配置

	计算节点	*登录节点*
操作系统	CentOS 7、McKernel	Red Hat Enterprise Linux7
编译器	GCC、Intel Compiler（C、C++、Fortran）	
MPI	Intel MPI、MVAPICH2	
库	Intel MKL LAPACK、FFTW、SuperLU、PETSc、METIS、Scotch、ScaLAPACK、GNU Scientific Library、NetCDF、Parallel netCDF、Xabclib、ppOpen-HPC、ppOpen-AT、MassiveThreads	
应用软件	mpijava、XcalableMP、OpenFOAM、ABINIT-MP、PHASE system、FrontFlow/blue、FrontISTR、REVOCAP、OpenMX、xTAPP、AkaiKKR、MODYLAS、ALPS、feram、GROMACS、BLAST、Rpackages, Bioconductor、BioPerl、BioRuby	
分布式FS		Globus Toolkit、Gfarm
任务调度	Fujitsu Technical Computing Suite	
调试程序	Allinea DDT	
剖析程序	Intel VTune Amplifier、Trace Analyzer & Collector	

（1）操作系统

登录时使用Red Hat Enterprise Linux系统，计算时使用Cent系统，未来可替换成McKernel系统。其中，Mckernel系统是日本理化学研究所正在开发的面向众核处理器的操作系统，预计该系统还将配置于“后京”系统中。

（2）编译器

Oakforest-PACS系统可使用GCC、Intel Compiler、XcalableMP等程序语言。其中的XcalableMP是理化学研究所与筑波大学正在共同开发的并行编程语言，该语言通过向用C语言或Fortran语言描述的代码添加指令字，使之更易于实现高性能的并行应用。

（3）应用软件

Oakforest-PACS系统主要使用OpenFOAM、ABINIT-MP、PHASE system、FrontFlow/blue等开源性应用软件。