E级计算之远景

邓越凡[1,2,3],**张黎浩**[1,2]

1.美国纽约州立石溪大学,美国纽约 11794
2.山东科学院国家超级计算济南中心,济南 250101
3.中山大学数据科学与计算机学院,广州 510006

摘要 超级计算机在当今科技发展中占有举足轻重的地位。在向着E级计算时代迈进之时,精确衡量超算的性能是一个事关超算架构和应用的关键问题。评价一台超算采用不同的基准会产生不同的结果。本文介绍超算中主要的3种排名及其对应的评测基准,并分析了超算本身的发展及应用远景。

关键词 超级计算机;戈登贝尔奖;Top500;Green500;Graph500;太湖之光;天河二号

20世纪60年代,Cray开发出的CDC 6600通常被认为是历史上第一台超级计算机。CDC 6600在标准数学运算中能够保持500 kflops(flops指每秒浮点运算数)的速度,是当时其他电脑的10倍。从那时开始,超级计算机速度大约每1.5年翻一番,非常接近著名的摩尔定律。而计算速度单位也从kflops,到Mflops(百万次浮点运算每秒),Gflops(十亿次浮点运算每秒),Tflops(万亿次浮点运算每秒)等。2016年中,由中国开发的最快超级计算机神威太湖之光的峰值速度达到125.4 Pflops(千万亿次浮点运算每秒),并能在广泛认可的LINPACK基准测试中保持93 Pflops的浮点运算速度。

大多数超级计算机是通过网络互连大量处理单元而构成的并行计算机。“处理器”代表单独的运算芯片,“核”代表芯片上接受并运行指令的基础运算单元,而“节点”代表计算机中一个或多个处理器组成的单元[1]。制造更快的超级计算机需要提高单个处理器的运算速度或者连接更多的处理器,或者两者同时进行。处理器速度的极快发展已带来处理器频率、访存、能源等一系列物理上难以逾越的限制。具体而言,提高单个处理器需要提高处理器时钟频率,或者提高每个时钟循环内的运算量,或者集合更多的处理核心。而每一种方法都会提高能耗,导致处理器温度过高,即使采用目前最好的冷却技术也难以保证半导体电路的正常运行。在单个处理器提速受限的情况下,为了提高超级计算机的运行速度,可行的方法是尽可能多的集成更多的处理核心来达到更高的总体速度。比如神威太湖之光拥有超过1000万的核。这项记录也必然会在不久的将来被正在开发中的新系统打破,下一个超算时代将会到达E级(百亿亿计算级别)[2]。以美国为例,IBM 的200 Pflops 的超算“Summit”预计将于2018年初在橡树岭国家实验室开始运行,而300 Pflops的超算“Sierra”预计将于2017年在劳伦斯·利弗摩尔国家实验室运行[3]。而以日本为例,继“京”之后,耗资9.1亿美元的1000 Pflops超算预计将于2020年发布。中国尚未正式发布下一个破纪录计划,但是不难猜测中国正在努力在这场激动人心的超算大赛中保持领先地位。

1 超级计算机的性能指标

1.1 Top500排行榜

几十年来,超算性能的定义等同于计算速度,以flops衡量。Top500 [4]以运行LINPACK基准测试所能达到的最高性能R_{max}(单位:Tflops)对500个超算系统进行排名。。排行榜同时提供很多有用的信息,包括制造商、地点、核数、网络互连技术等。表1为2016年6月排行榜前10的超算,其中R_{peak}(单位:Tflops)代表理论峰值速度。

10台超算中,中国的超级计算机位居前两位。其中2013年7月发布的天河二号,已经在此之前连续6次以33.86 Pflops的运行速度排行榜首。其他4台在美国,日本、瑞士、德国和沙特阿拉伯各有1台。

1.2 Green500排行榜

在持续几十年的运行速度的竞赛中,建造和能耗的预算并不在考虑之中,但从业者逐渐发现超级计算机正面临着能耗过高的限制。2007年,侧重于超算能效的Green500[5]排行榜开始发布。用电效率Mflops/W,即每W功率可以支持多少 Mflops的运行速度。最近,Green500和Top500宣布合并使用同样的提交规则来标准化能耗测量标准。相关的说明文档详细规定了能耗测量所需要考量的因素,并设定了由低到

收稿日期:2016-08-04;修回日期:2016-08-25
作者简介:邓越凡,教授,研究方向为应用数学、计算科学,电子信箱:yuefan.deng@stonybrook.edu
引用格式:邓越凡,张黎浩.E级计算之远景[J].科技导报,2016,34(21):85-94;doi:10.3981/j.issn.1000-7857.2016.21.012

表1 2016年6月Top500榜单前十位的系统

Table 1 Top 10 systems of June 2016 Top500 list

地点	超算系统	核数	R_{max}/Tflops	R_{peak}/Tflops
国家超算无锡中心 中国	Sunway TaihuLight	10649600	93014.6	125435.9
国家超算广州中心 中国	Tianhe-2	3120000	33862.7	54902.4
橡树岭国家实验室 美国	Titan	560640	17590.0	27112.5
劳伦斯利弗莫尔国家实验室 美国	Sequoia	1572864	17173.2	20132.7
理化学研究所 日本	K computer	705024	10510.0	11280.4
阿贡国家实验室 美国	Mira	786432	8586.6	10066.3
洛斯阿拉莫斯国家实验室 美国	Trinity	301056	8100.9	11078.9
瑞士国家超算中心 瑞士	Piz Daint	115984	6271.0	7788.9
斯图加特高性能计算中心 德国	Hazel Hen	185088	5640.0	7403.5
阿卜杜拉国王科技大学 沙特阿拉伯	Shaheen II	196608	5537.0	7235.2

高3种测量品质。这对提交的数据提出了更高的要求,以保证最后能效排名的准确性。同时Green500和Top500依然是不同网站上独立的两个排行榜。

表2为2016年1月Green500榜单前10位系统,出人意料的是,Green500上大多数高排名的超算没有在Top500的前列出现。通常节能型的超级计算机是在给定的能耗限制下仔细设计建造的,以求达到可能的最高能效。通常这些机器规模较小,能耗只有几万瓦。制造同时拥有顶尖计算速度和高能效的大规模超级计算机依然是一项具有挑战的任务。

1.3 Graph500排行榜

Top500排行榜采用LINPACK基准测试超级计算机在解稠密线性方程组时的性能。然而对于超算系统在包括数据密集型应用在内的许多其他应用中,Top500并没有提供有用的信息。2010年,一个小组开始着手研究大数据应用方面的新的性能基准,并在当年发布了Graph500[6]。该基准用于衡量超算通信子系统的性能,它测量的是在一个大型无向图上执行广度优先算法时,每秒遍历边缘的数量,单位为Gteps(每秒10亿遍历边缘数)。

该基准包括一个可扩展的数据生成器,可以生成包含所有边起点和终点边的数组。第一个核心进程生成一个无向图,其格式能够被接下来所有的核心进程所用。此后不允许任何改动,以防止某些核心进程会因此获益。第二个核心进程则是对生成的图执行广度优先算法。两个进程都进行计时。根据输入规格大小分成6个问题类型:从最小10^{10}字节的"toy"到10^{15}字节的"huge"。

表2 2016年6月Green500榜单前10位的系统

Table 2 Top 10 systems of June 2016 Green500 list

地点	超算系统	能效/(Gflops·W^{-1})	Top500排名
Advanced Center for Computing and Communication, RIKEN	ZettaScaler-1.6, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband FDR, PEZY-SCnp	6673.8	94
Computational Astrophysics Laboratory, RIKEN	ZettaScaler-1.6, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband FDR, PEZY-SCnp	6195.2	486
National Supercomputing Center in Wuxi	Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway	6051.3	1
GSI Helmholtz Center	ASUS ESC4000 FDR/G2S, Intel Xeon E5-2690v2 10C 3GHz, Infiniband FDR, AMD FirePro S9150	5272.1	440
Institute of Modern Physics (IMP), Chinese Academy of Sciences	Sugon Cluster W780I, Xeon E5-2640v3 8C 2.6GHz, Infiniband QDR, NVIDIA Tesla K80	4778.5	446
Stanford Research Computing Center	Cray CS-Storm, Intel Xeon E5-2680v2 10C 2.8GHz, Infiniband FDR, Nvidia K80	4112.1	122
Internet Service	Inspur TS10000 HPC Server, Intel Xeon E5-2620v2 6C 2.1GHz, 10G Ethernet, NVIDIA Tesla K40	3775.5	305
Internet Service	Inspur TS10000 HPC Server, Intel Xeon E5-2620v2 6C 2.1GHz, 10G Ethernet, NVIDIA Tesla K40	3775.5	306
Internet Service	Inspur TS10000 HPC Server, Intel Xeon E5-2620v2 6C 2.1GHz, 10G Ethernet, NVIDIA Tesla K40	3775.5	307
Internet Service	Inspur TS10000 HPC Server, Intel Xeon E5-2620v2 6C 2.1GHz, 10G Ethernet, NVIDIA Tesla K40	3775.5	308

Graph500还是一个非常年轻的项目。最新的排行榜只列出了211台超算,离真正500有一段距离。大多数Graph500的小规模超算并没有出现在Top500上。大约70%的Graph500超算系统来自美国和日本(表3)。

表3　2016年6月Graph500榜单前10位系统

Table 3　Top 10 systems of June 2016 Graph500 list

排名	系统名称	节点数	核数	问题规格	Gteps
1	K computer	82944	663552	40	38621
2	Sunway TaihuLight	40768	10599680	40	23756
3	Sequoia	98304	1572864	41	23751
4	Mira	49152	786432	40	14982
5	Juqueen	16384	262144	38	5848
6	Fermi	8192	131072	37	2567
7	Tianhe-2	8192	196608	36	2061
8	Turing	4096	65636	36	1427
9	Blue Joule	4096	65636	36	1427
10	DIRAC	4096	65636	36	1427

2　2016年超算分析

对2016年6月Top500的超算系统进行分析。Green500列出了Top500所有超算的能效,使得Top500不完整的能效数据得到补充。而作为一个新排行榜,Graph500以一个非常不同的角度评测超算,因此Top500的超算系统在Graph500上出现得不多。同时Top500上很多超算系统也是专门对LINPACK基准做过优化的,所以在分析中没有考虑Graph500的排名。

2.1　性能散点图

与文献[7]、[8]类似,用一幅能效相对于LINPACK效率的散点图(图1)检视各超算系统。图1中空心圈或者实心圆的圆心代表相应的LINPACK和能效,圈和圆的面积与持续运算速度R_{max}成正比,圈代表超算系统没有加速器,实心圆代表有加速器,颜色用于区分网络互连类型。

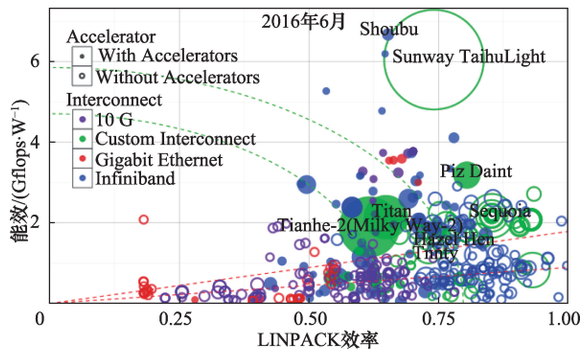

图1　2016年6月Top500和Green500的LINPACK效率和能效散点图

Fig. 1　Scatter plot for LINPACK and power efficiencies from June 2016 Top500 and Green500

为了简化,将私有互连和自定义互连合并为一类。因此,在一般散点图通常两个维度(LINPACK效率和能效)的基础上,额外增加了持续速度、互连类型、是否有加速器3个维度。图1中绿色的曲线把散点图分为3个区域,每个区域包含了1/3的点。红色射线和绿色曲线类似,将把散点图分成3个有同样点数的区域。最理想的超算系统毫无疑问拥有很高的能效和LINPACK效率,同时能达到很高的持续计算速度,这些系统会以大面积的圈或者圆的形式出现在散点图的右上角。

需要指出的是,LINPACK效率并不是一个万能的性能指标,因为它只能反映一个超算系统在解稠密线性方程组的性能表现。很明显,很多现实中的应用会用到其他的核心功能,包括快速傅里叶变换、非线性微分方程组、全局优化问题等。用LINPACK效率作为评判超算系统在这些应用上的性能表现显然容易造成误导。

图1散点图标示了2016年6月Top500的超算系统。从图1中可以看出,神威太湖之光不仅拥有最高的运行速度,同时拥有很高的LINPACK效率和能效。尽管事实上大部分的超算系统用InfiniBand作为网络互连,但是顶尖的系统更倾向于用自定义和私有的互连技术。另外容易发现,使用同一种互连技术的超算系统,使用加速器或者协处理器的LINPACK效率偏低。这个现象和通常观点一致:使用加速器的超算系统难以在包括LINPACK的很多应用上获得高性能表现。

2.2　性能和架构细节

2.2.1　处理单元

Intel提供的处理器所占比最高,大约89%。在Top500中,Intel Xeon E5是最受欢迎的处理器系列。Intel Xeon Phi协处理器同时也被用于补充加强Xeon E5系列的性能和能耗表现(图2)。在前10的超算系统中,有5台使用了Intel处理器。

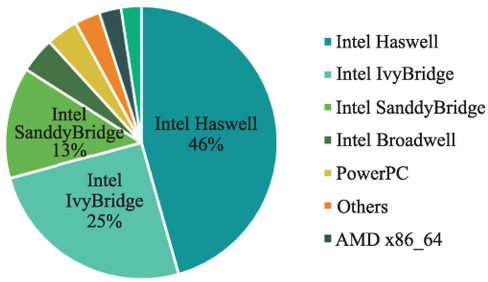

图2　2016年6月Top500处理器份额

Fig. 2　Processor generation system shares from June 2016 Top500

而加速器或协处理器方面,Top500共有94台超算使用了该项技术,相比2015年11月的104台略有减少,同时GPU加速器最为常用(图3)。NVIDIA的Telsa GPU在加速器市场占据主导地位,Intel在该市场的巨大努力尚未产生明显的影

响。从Top500的份额可以看出,67台系统采用了NVIDIA GPU,Xeon Phi只被用于26台,而有3台系统同时使用了NIVIDIA和Intel Xeon Phi。其他两种小众GPU也有被用到:3台系统使用了ATI Radeon,2台使用了PEZY技术。加速器市场Intel和NVIDIA的竞争正日趋激烈。Intel最新的Xeon Phi产品线,Knights Landing极大吸引了业界注意。Intel决心在2016年投放10万芯片到高性能计算市场,以增加其在市场和性能方面的所占份额。

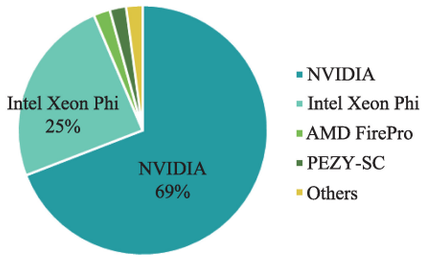

图3　2016年6月Top500加速器/协加速器份额

Fig. 3　Accelerator/Co-Processor system share from June 2016 Top500

神威太湖之光使用了由上海高性能集成电路设计中心制造的中国产众核处理器SW26010。SW26010是针对高程度并行计算设计的,包含260个独立的处理核心。它是一个整体的芯片,由4个核组组成,每个核组包含1个管理处理单元MPE和64个运算处理单元CPE[9]。

通常而言,Top500的异构众核系统是通过将多核CPU和众核加速器(比如Intel Xeon Phi或NVIDIA GPU)组合建造而成的。例如,之前连续排名第一的天河二号包含16000个节点,每个节点包含2个Intel Xeon Ivy Bridge处理器和3个Xeon Phi协处理器,共3120000个计算核心。这样的设计会在CPU和GPU内存之间产生显著的数据传输开销[10],从而影响很多应用的并行效率。由图1可以看出,在其他条件相同的情况下,有加速器的超算系统一般LINPACK效率会更低。太湖之光采用的SW26010处理器消除了通过PCIe在CPU和加速器内存之间传输的影响,因而其LINPACK效率大大优于天河二号。

2016年8月份,富士通宣布其扩展了ARM处理器,将向量指令集包含其中以推进其64位V8架构。ARM处理器相比x86系列有着更高的能效。同时,其可伸缩向量扩展指令集(scalable vector extensions,SVE)支持从128~2048位的向量长度,具有极大的灵活性,大大扩展软硬件及应用开发者的发挥空间。SVE是一种负载/存储架构,采用最多32个向量缓存器,提供了一系列针对科学运算的新指令集。富士通打算采用SVE使得京的后续版本(post-K)的系统性能比前一代提升50倍,能效提升15倍。

2.2.2　网络互连系统

Top500主要使用了InfiniBand、以太网(包括GbE和10GbE)和自定义互连3种网络互连技术。由图4可以看出,高排名的超算系统更倾向于用自定义的互连技术。InfiniBand和以太网都被广泛采用,然而考虑LINPACK效率的话InfiniBand明显更胜一筹。旧的千兆以太网(GbE)的性能表现最糟糕,它们集中在在散点图的左下角,代表着低LINPACK效率和低能效。这些系统虽然制造成本最低,但是运行开销很高,处理速度低而能耗又高,所以并不理想。下一代的10GbE的性能表现更好,但是LINPACK效率依旧低于InfiniBand。InfiniBand系统的平均LINPACK效率是77.3%,而GbE的只有42.7%。

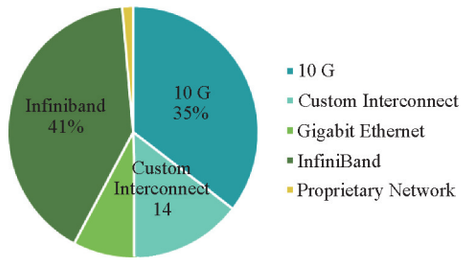

图4　2016年6月网络互连类型份额

Fig. 4　Interconnect family shares from June 2016 Top500

2.2.3　操作系统

作为Unix操作系统的变体,Linux在超算系统中占有绝对的主导地位。Top500有497个系统采用了Linux,只有3个使用了Unix。

2.2.4　并行构架

SMP于2003年从Top500上消失之后,大规模并行机(MPP)和集群(cluster)成为了两种主流并行架构。如果使用对称多处理机(SMP)设计,增加处理器会使得运行效率不断下降,因为大量的处理器同时访问同一个内存会造成访存冲突。由图5可以看出,cluster是主要被采用的架构,占Top500中的86%。然而,前10中只有1/5采用cluster,在前50中,cluster和MPP所占份额相差无几。从LINPACK效率来看,MPP要明显优于cluster。

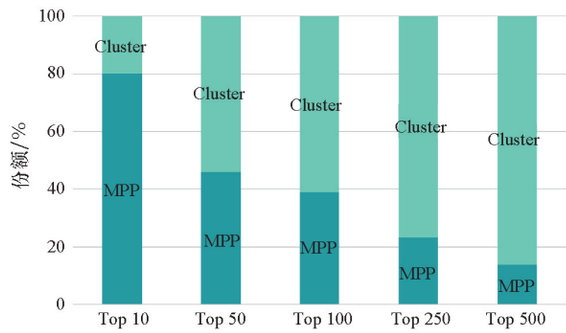

图5　2016年6月Top500不同排名区间并行架构份额

Fig. 5　Architecture system shares in different rank range from June 2016 Top500

图6展示了另一个有趣的现象:MPP通常被那些使用自定义或者私有网络互连的超算所采用,而cluster基本被那些使用以太网或者InfiniBand的系统采用。在所有204个使用InfiniBand的系统中,只有一个采用了MPP。这样的选择并不是偶然,因为在一个cluster系统中,每个节点包含自己的内存,硬盘和网络接口,这样的话节点能够被标准商用网络通信技术(比如InfiniBand和GbE)简单连接起来。

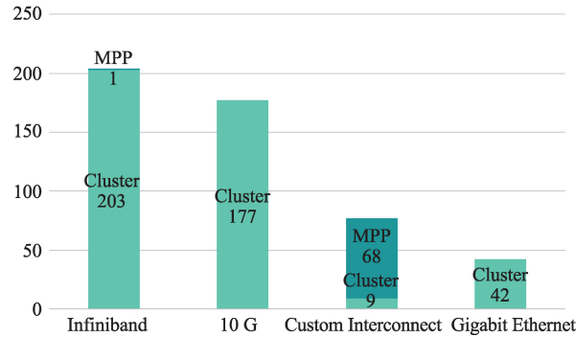

图6 2016年6月Top500不同网络互连系统使用的并行架构

Fig. 6 Parallel architectures used by systems with different interconnect family from June 2016 Top500

3 2016年之前的超算分析

如图7所示,摩尔定律展现了过去60年超算速度的发展。如今的智能手机已经可以达到1~10 Gflops的运行速度[11],超过了20年前那些超大规模的超级计算机。表4列出了从1964年至今最快超算的发展变化。在20世纪60—90年代,Cray持续刷新最快超算的纪录,基本上垄断了超算市场。美国和日本之间在Top500的竞争愈演愈烈时,中国于2001年也加入了角逐。经过了飞速发展,终于在2016年6月凭借167台超算系统在Top500上与美国居于同样重要的地位。事实上,从2013年开始,中国已经连续7次登上排名第一的宝座。

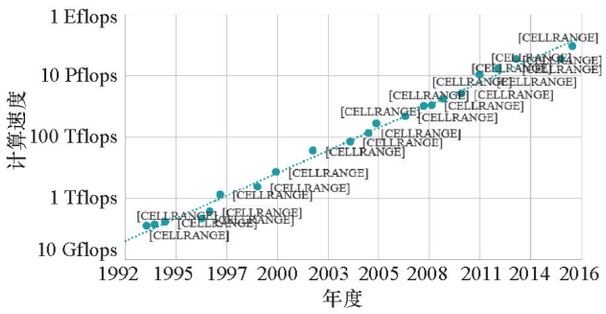

图7 1993年以来的超算计算速度的发展

Fig. 7 Speeds of the fastest supercomputers since 1993

包括Intel在内的半导体芯片制造商从2015年就开始注意到,从2012年22 nm到2016年14 nm开始,单个处理器的提速开始放缓。但在超算领域中,更多的处理器被装入更大的系统用来弥补单个处理器的速度放缓,最终使得摩尔定律依然持续下去。

表4 1964年至今的最快超算

Table 4 Fastest supercomputers since 1964

年份	超算系统	峰值速度	国家
1964—1968	CDC6600	1 Mflops	美国
1969—1975	CDC7600	10 Mflops	美国
1976—1982	Cray-1	146 Mflops	美国
1983—1985	Cray X-MP/4	713 Mflops	美国
1985—1987	Cray-2	1.95 Gflops	美国
1988—1989	Cray Y-MP/832	2.14 Gflops	美国
1990—1991	Fujitsu VP2600/10	4 Gflops	日本
1992	NEC SX-3/44	20 Gflops	日本
1993	Thinking Machines CM-5/1024	59.7 Gflops	美国
	Numerical Wind Tunnel	124 Gflops	日本
1994	Intel Paragon XP/S140	143 Gflops	美国
1994—1995	Numerical Wind Tunnel	170 Gflops	日本
1996	Hitachi SR2201	368 Gflops	日本
1997—2000	ASCI Red	2.38 Tflops	美国
2001	ASCI White	4.94 Tflops	美国
2002—2004	Earth Simulator	36.9 Tflops	日本
2004—2007	IBM Blue Gene/L	70.8 Tflops	美国
		136 Tflops	
		281 Tflops	
		478 Tflops	
2008	IBM Roadrunner	1.03 Pflops	美国
2009	Jaguar	1.75 Pflops	美国
2010	Tianhe-1A	2.57 Pflops	中国
2011	Fujitsu K computer	10.5 Pflops	日本
2012	IBM Sequoia	16.3 Pflops	美国
	Cray Titan	17.6 Pflops	美国
2013—2015	Tianhe-2	33.9 Pflops	中国
2016	Sunway TaihuLight	93 Pflops	中国

3.1 性能散点图

与图1类似,生成了2013、2010以及2007年的性能散点图(图8~图10)。

在图8~图10中,x轴相同,而y轴根据当年最高能效调

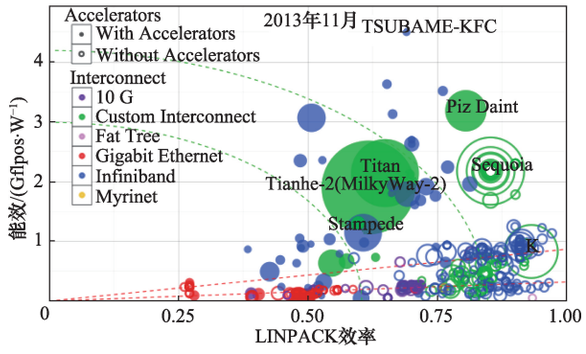

图8 2013年11月Top500,LINPACK效率和能效散点图

Fig. 8 Scatter plot for LINPACK and power efficiencies from November 2013 Top500 list

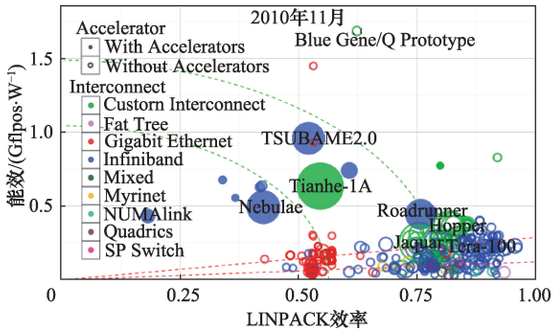

图9　2010年11月Top500,LINPACK效率和能效散点图

Fig. 9　Scatter plot for LINPACK and power efficiencies from November 2010 Top500 list

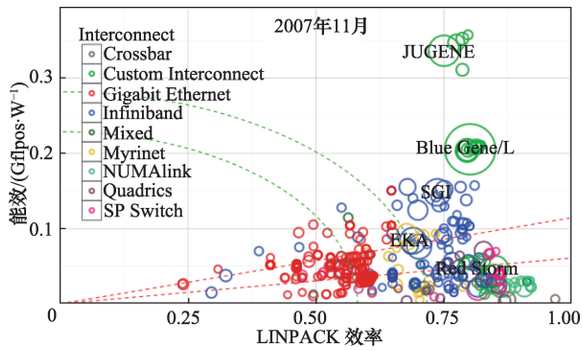

图10　2007年11月Top500,LINPACK效率和能效散点图

Fig. 10　Scatter plot for LINPACK and power efficiencies from November 2007 Top500 list

整,包含了过去10年超算发展的丰富信息。根据分析发现:1）排名第1和第500之间的差距在拉大,也就是说最快的系统的提速幅度更大。2）最高能效从2007年的0.357 Gflops·W^{-1},到2010年的1.684 Gflops·W^{-1},到2013年的4.503 Gflops·W^{-1},最后到2016年的6.674 Gflops·W^{-1}。这个走势似乎也符合摩尔定律,每3年增加4倍。3）除了神威太湖之光达到了6.051 Gflops·W^{-1},其他的Top500高排名系统在Green500表现并不出彩。因此神威太湖之光同时是最快,也是最节能的超算系统。4）使用GbE的超算系统的能效和LINPACK效率最低。这类系统的数量在逐年减少,同时它的下一代标准10GbE从2007开始被越来越多的系统所采用。5）整体上看使用加速器的系统能效更高,但是LINPACK效率变低。

3.2　发展细节

基于GbE的系统很长时间内占据着统治地位,从2005—2011年保持着Top500上超过40%的比例。但是从2009年开始这个比例有下降的趋势,到现在只有8.4%,也不难猜想在接下来几年内GbE会逐渐从Top500上消失。它的升级版本10GbE近年来保持着稳定的增长,从2010年的2台到2016年的177台。目前还没有迹象显示10GbE的热门程度会在接下来的2~3年内减弱。而Infiniband一直都被广泛采用,是2012—2016年Top500使用最多的网络互连技术(图11)。设计创新型网络互连的竞争日趋激烈。

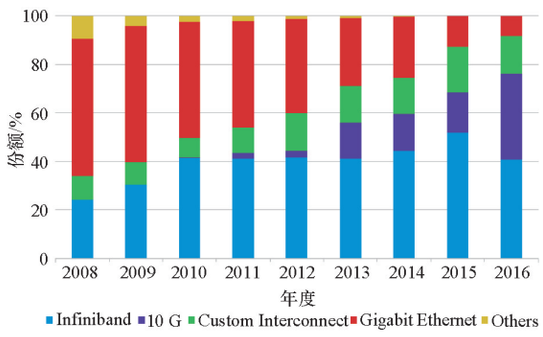

图11　Top500中网络互连类型的份额变化

Fig. 11　Time-varying shares of the interconnect families on Top500

Linux在过去10年一直都占据主导地位,并且不断加强。由图12可以看出。在这期间,Unix从2007年的60台到2016年的3台,接近消失。

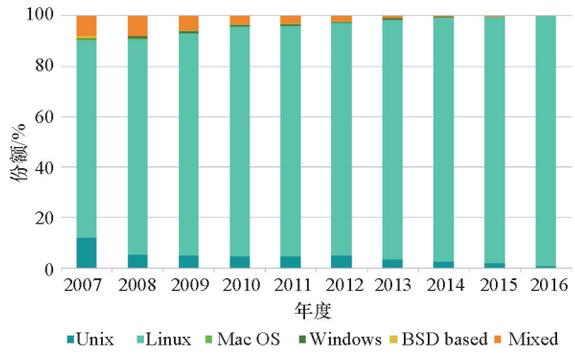

图12　Top500中操作系统的份额变化

Fig. 12　Time-varying shares of the operating systems on Top500

给超算系统加上加速器还是较新的技术,2007年的Top500上仅有一台机器使用了加速器。使用加速器和协处理器的系统数量在逐年增加,2015年11月增加到103台,然而在今年稍稍减少到94台。这也意味着加速器方法正在走过它的黄金期,对于这些使用众核加速器的异构众核系统,LINPACK效率总显得要低一些。

中国的超算发展在过去10年经历了突破性的进展。自从天河二号第一次在2013年获得Top500排名第一,最快超算的皇冠再也没有离开中国。相较2015年底的109台,2016年中国有167台超算系统上榜。2016年6月也是中国第一次在Top500的数量上超过了美国的165台,更远超于第三名日本的29台(图13)。另外,中国也是超算性能占比最高,这归功于排名前两名的超算:天河二号和太湖之光。

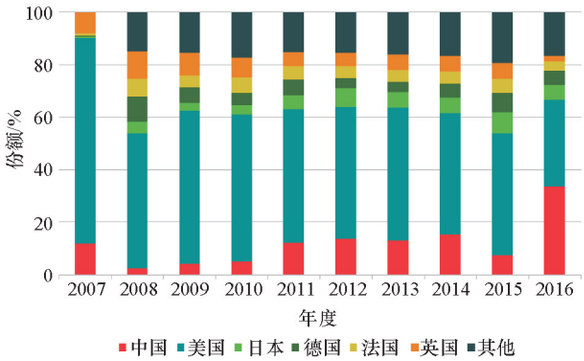

图13　Top500中不同国家所占份额的变化

Fig. 13　Time-varying shares of countries on the Top500 list

4　超算应用

4.1　戈登贝尔奖

由国际计算机学会(ACM)管理,每年的超级计算机大会上ACM 戈登贝尔奖用来表彰那些在高性能计算方面的杰出成就。戈登贝尔奖纪录了并行计算的发展,强调表彰那些在科学工程和大规模数据分析方面的高性能计算应用创新。戈登贝尔奖可授于在重大的科学工程问题上获得的峰值性能、可扩展性以及求解时间方面的成就。表5列出了2006年来所有的获奖研究、小组及其成就。目前而言,中国尚未有项目获得戈登贝尔奖,相信最近不断的努力很快会获得认可。中国的3个项目:钛合金微结构演化相场模拟、千万亿次850万核可扩展非静力大气动力全隐求解器、MASNUM海浪模式的移植与优化,已经被2016年戈登贝尔奖提名[12]。

在科学工程问题中,分子动力、生物、航空和地理等方面的应用至关重要,而模拟问题的规模是成功解决问题的重要基准(图14、图15)[13-14]。获奖要求开发者引入算法、模型的创新,并且并行程序上获得显著的性能表现和可扩展性。2015年,德州大学Austin分校、IBM研究组、纽约大学和加州技术研究所开发了创新性的算法来模拟地壳运动,并且同时加入模拟地幔运动的影响。该研究小组的代码达到了前所未有的97%的并行效率,并且扩展到Sequoia超算的160万的核心上。2013年,ETH Zurich和IBM的科学家完成了13万亿单元的流体动态模拟,达到了14.4 Pflops的持续性能表现——Sequoia的73%的理论峰值速度。

Top500基准的开发者之一Jack Dongarra提到,太湖之光3个被提名戈登贝尔奖的应用中,两个已经达到了30~40 Pflops的持续运行速度,超过了天河二号的LINPACK最大性能表现。最高持续速度达到太湖之光31.6%的理论峰值速度[15]。

表5　2006—2015年所颁发的戈登贝尔奖

Table 5　Gordon Bell prize, 2006-2015

年份	获奖论文	研究组	成果
2015	An extreme-scale implicit solver for complex PDEs: Highly heterogeneous flow in earth's mantle	The University of Texas at Austin	可扩展性
2014	Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer	D.E. Shaw Research	性能表现
2013	11 PFLOP/s simulations of cloud cavitation collapse	ETH Zürich, Switzerland	峰值性能表现
2012	4.45 Pflops astrophysical *N*-body simulation on K computer: the gravitational trillion-body problem	University of Tsukuba	可扩展性和耗时
2011	First-principles calculations of electron states of a silicon nanowire with 100,000 atoms on the K computer	Next-Generation Supercomputer R&D Center	持续性能表现
	Peta-scale phase-field simulation for dendritic solidification on the TSUBAME 2.0 supercomputer	Tokyo Institute of Technology	可扩展性和耗时
2010	Petascale direct numerical simulation of blood flow on 200 K Cores and Heterogeneous Architectures	Georgia Institute of Technology	性能表现
	42 TFlops hierarchical *N*-body simulations on GPUs with applications in both astrophysics and turbulence	Nagasaki University	价格/性能表现
2009	A scalable method for *ab initio* computation of free energies in nanoscale systems	Oak Ridge National Laboratory	峰值性能表现
	Millisecond-scale molecular dynamics simulations on Anton	D. E. Shaw Research	特殊范围
2008	New algorithm to enable 400+ Tflop/s sustained performance in simulations of disorder effects in high-T_c superconductors	Oak Ridge National Laboratory	峰值性能表现
	Linearly scaling 3D fragment method for large-scale electronic structure calculations	Lawrence Berkeley National Laboratory	算法创新
2007	Extending stability beyond CPU millennium: a micron-scale atomistic simulation of Kelvin-Helmholtz instability	Lawrence Livermore National Laboratory	性能表现
2006	Large-scale electronic structure calculations of high-Z metals on the BlueGene/L platform	University of California, Davis	峰值性能表现
	The BlueGene/L supercomputer and quantum Chromo Dynamics	IBM T.J. Watson Research Laboratory	特殊贡献

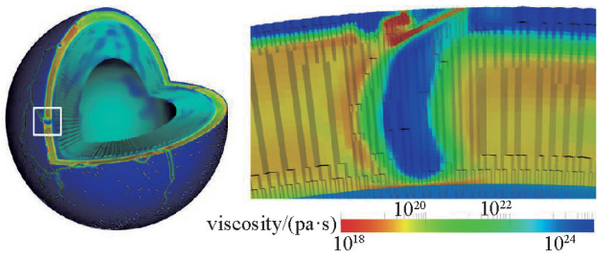

图14　太平洋板块中一块俯冲板块横截面，展示了非线性地幔流模拟中的地幔凝滞性，荣获2015年戈登贝尔奖

Fig. 14　Pacific plate in a cross section of the subducting plate, showing the nonlinear mantle flow stagnation in the simulation, won the Gordon Bell prize in 2015

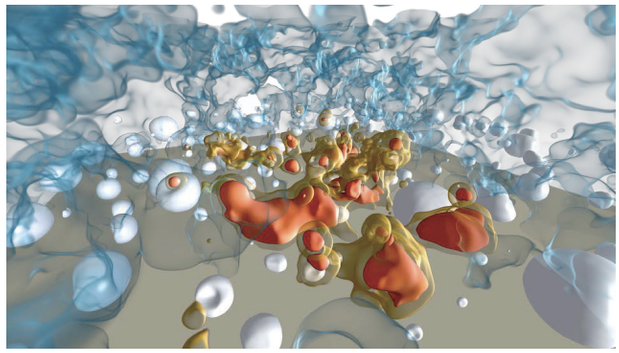

图15　压力/界面的体积/等值面的展示(白色等值面代表气泡，橙色/绿色代表高/低压区域，荣获2013年戈登贝尔奖)

Fig. 15　Pressure/interface volume / ISO surface display

4.2　大数据

如今数据量正在急剧膨胀，一部分原因是如今大量的移动传感器设施、软件日志、摄像装置、手机和无线网络使得数据收集的成本大为降低。数据是如此庞大而复杂，传统的数据处理方法难以派上用场，数据存储和访问的架构创新亟待进行。2004年，Google发表了关于MapReduce进程的文章，这种进程提供了一种并行处理模型，相关的实践也得以发布用来处理大型数据。在MapReduce的Mapping过程中，查询数据任务被拆分并分配到并行节点中处理。然后在Reduce过程中，处理结果被收集并发送出去。MapReduce框架被实际应用到一个叫作Hadoop的Apache开源项目中，并获得了多种编程语言的支持(图16)。

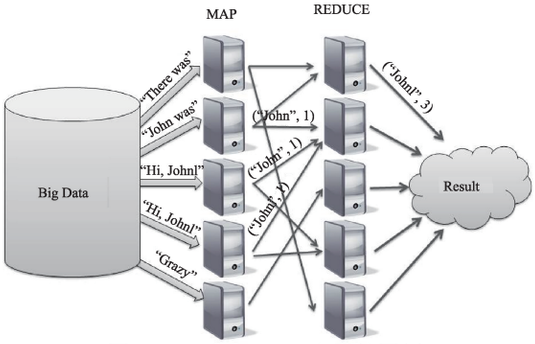

图16　MapReduce应用的图例

Fig. 16　The application of MapReduce legend

(来源：http://dme.rwth-aachen.de/de/research/projects/mapreduce)

除了并行架构外，制造商同时也在优化他们的系统，加速大数据应用。IBM重新设计他们下一代的计算机，最大程度降低存储和处理器之间的数据访存[16]。2014年，IBM获得了3.25亿美元的资金，在2017年前完成两台新的超算："Summit"和"Sierra"[17]。两台超算都依赖于IBM提出的"以数据为中心"的架构，将计算能力分配于整个系统，在不同的点处理数据，而不是在系统内将数据来回传输。

4.3　材料科学

工程材料，比如家用于芯片制造的硅和光纤里的玻璃，是现代世界的基础。然而设计新材料在历史上从来都需要很多令人沮丧而低效的猜测工作。超算和量子力学的应用，使得科学家能够从原子开始设计新材料，而不用事先进行传统的实验。

材料科学正在经历新的革命。计算机处理性能的发展，同时还有20世纪60和70年代间Walter Kohn和John Pople对量子力学方程求解问题上不失精度的简化工作，使得运用超算和基础物理从零开始设计新材料变为可能[18]。

图17显示了文献[19]胶状凝胶模拟模拟时的一幅快照。应用超级计算机，研究人员能够一次同时模拟包含了75万个分子的体系，从而更好的研究胶体复杂结构的重组过程。

4.4　数值天气预报

高精度的数值天气预报是高性能计算中的一个最具挑战性的应用，需要在细密网格上进行快速高精度的模拟。预报天气需要用到一系列不同的模拟和建模技术。典型的工作量包含数据同化、决定性预测模型和集合预测模型。另

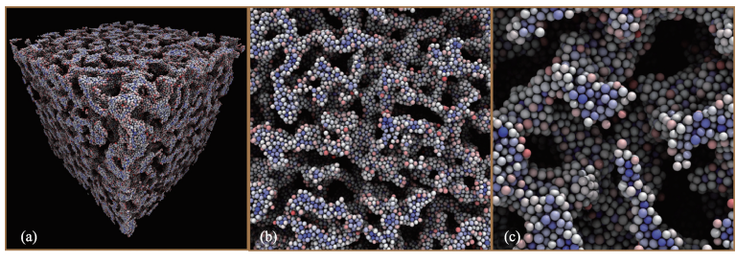

图17　来自文献[19]的胶体模拟快照

Fig. 17　Colloid simulation snapshot from literature [19]

外,一些专门的模型可能会应用于其他方面,比如极端天气、空气质量、洋流和道路情况等,图18为使用437 GPUs模拟经过日本的台风[20]。为了提高预测能力,处理大量数据并进行复杂计算至关重要,因此需要那些性能强劲的超级计算机。

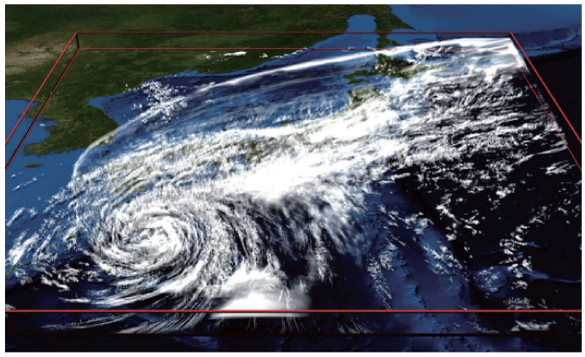

图18 使用437 GPUs模拟经过日本的台风

Fig. 18 Using 437 GPUs simulation after a typhoon in Japan

5 结论与展望

尽管在硬件、编程模型、软件方面有着诸多障碍,E级次超级计算机也会在2020年之前被开发出来。如今,美国、中国、日本以及印度、俄罗斯和欧盟都加入了开发E级次超算的角逐。2015年,美国总统奥巴马签署了一项决策性法令,授权建立名为NSCI的新的超级计算研究机构。其目标是完成第一台E级次超级计算机的开发,并且运行速度是太湖之光的10倍。

当然,让应用能够完全挖掘E级次超算的能力和建造超算一样,都是富有挑战性的工作。主要的问题在于能耗的急剧增加、提供数据局部性、保证高并行性和可靠性。

同时,其他可能的未来超算也正在开发中。尽管听起来像是科幻中的概念,比如量子超算和生物超算,最近几年也正在逐渐展现出来。D-Wave 2X量子计算机[21]就是这类运用量子力学计算机中的一个佼佼者,其大小相当于一个大冰箱,远小于性能类似的传统计算机。 2015年,Google的一个研究小组和NASA工程师发表声明[22],D-Wave计算机能够以普通单核处理器1亿倍的速度解出一个最优化问题。

欧盟资助的ABACUS项目的研究人员已经制造出了一台生物超级计算机的模型机,而且能够在持续运行的同时保持高能效[23]。该模型机是由ATP——人体细胞中用于供能的物质进行运作的。这台生物超算可以像传统超算一样运行并行计算,并且成功高效地解决复杂的数学问题。不过该项目小组也承认从模型机到完整规格的生物超算还有很长的一段路要走。

在向着E级次计算时代迈进时,系统架构和编程模型需要做诸多改进面对各种新的挑战。 近些年超级计算机计算速度的提升似乎有所放缓,有许多人怀疑摩尔定律是否还能持续,但神威太湖之光令人瞩目的登场向世界宣告摩尔定律依旧有效。事实上,人们对更高计算速度的渴望也永远不会消退。

致谢:本文撰写过程中,得到中国科学院院士郝柏林的支持。

参考文献(References)

[1] Deng Y, Zhang P. Perspectives of exascale computing[J/OL]. Journal of New Computing Architectures and Applications, [2016-08-15], 1. http://www.ams.stonybrook.edu/~penzhang/papers/Perspective- 2010_DENG.pdf.

[2] Mitropol′skii Y I. Electronic components and architecture of future supercomputers[J]. Russian Microelectronics, 2015, 44: 139-153.

[3] Nvidia News. U.S. to Build Two Flagship Supercomputers for National Labs[EB/OL]. [2016-9-22]. http://nvidianews.nvidia.com/news/u-s-to-build-two-flagship-supercomputers-for-national-labs.

[4] Top500. TOP500 Supercomputers[EB/OL]. [2016-09-22]. http://www.top500.org.

[5] Green500. Green500 Supercomputers[EB/OL]. [2016-09-22]. http://www.green500.org.

[6] Graph500. Graph500 Supercomputers[EB/OL]. [2016-09-22]. http://www.graph500.org.

[7] Deng Y, Zhang P, Marques C, et al. Analysis of Linpack and power efficiencies of the world′s TOP500 supercomputers[J]. Parallel Computing, 2013, 39(6): 271-279.

[8] Deng Y, Korobka A, Lou Z, et al. Perspectives on petascale processing[J/OL]. [2016-09-22]. http://www.ams.sunysb.edu/~penzhang/papers/Perspective-2008_DENG.pdf.

[9] Zheng F, Xu Y, Li H, et al. A homegrown many-core processor architecture for high-performance computing[J]. Scientia Sinica Informationis, 2015, 45(4): 523-534.

[10] Daga M, Aji A M, Feng W. On the efficacy of a fused CPU+ GPU processor (or APU) for parallel computing[C]//2011 Symposium on Application Accelerators in High-Performance Computing. Knoxville: IEEE, 2011: 141-149.

[11] Hurd M. Top500 - is that a supercomputer in your pocket?[EB/OL]. [2016-08-25]. http://meanderful.blogspot.com/2015/07/top500-is-that-supercomputer-in-your.html.

[12] National Supercomputing Center in Wuxi.Application domains[EB/OL]. [2016-08-25]. http://www.nsccwx.cn/wxcyw/case.php?i=58&word=case.

[13] Rudi J, Malossi A C I, Isaac T, et al. An extreme-scale implicit solver for complex PDEs: highly heterogeneous flow in earth′s mantle[C]//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. New York: ACM, 2015. doi: HYPERLINK "http://dx.doi.org/10.1145/2807591.2807675" \t "_self" 10.1145/2807591.2807675 .

[14] Rossinelli D, Hejazialhosseini B, Hadjidoukas P, et al. 11 Pflop/s simulations of cloud cavitation collapse[C]//International Conference on High PERFORMANCE Computing, Networking, Storage and Analysis. IEEE, 2013: 1-13.

[15] Ceder G, Persson K. How supercomputers will yield a golden age of materials science[J/OL]. Scientific American, [2016-08-25]. http://www.scientificamerican.com/article/how-supercomputers-will-yield-a-golden-age-of-materials-science/.

[16] Zia R N, Landrum B J, Russel W B. A micro-mechanical study of coarsening and rheology of colloidal gels: Cage building, cage hopping, and Smoluchowski´s ratchet[J]. Journal of Rheology (1978-present), 2014, 58 (5): 1121-1157.
[17] Shimokawabe T, Aoki T, Ishida J, et al. 145 Tflops performance on 3990 GPUs of TSUBAME 2.0 supercomputer for an operational weather prediction[J]. Procedia Computer Science, 2011, 4: 1535-1544.
[18] Navarro A."Google´s D-wave 2X quantum computer 100 million times faster than regular computer chip[EB/OL]. [2016-08-25]. http://www.techtimes.com/articles/114614/20151209/googles- d- wave- 2x- quantum-computer-100-million-times-faster-than-regular-computer-chip.htm.
[19] Neven H. When can quantum annealing win?[EB/OL]. [2016-08-25]. https://research.googleblog.com/2015/12/when- can- quantum- annealing-win.html.
[20] CORDIS. Researchers have created a breakthrough model biological supercomputer[EB/OL]. [2016-08-15]. http://phys.org/news/2016-03-breakthrough-biological-supercomputer.html.
[21] Hsu J. IBM is redesigning supercomputers to solve big data problems[EB/OL]. [2016-09-22]. http://spectrum.ieee.org/tech-talk/computing/hardware/ibm-redesigned-supercomputers-to-solve-big-data-problems.
[22] Dongarra J. Report on the sunway TaihuLight System[EB/OL]. [2016-9-22]. http://www.netlib.org/utk/people/JackDongarra/PAPERS/sunway-report-2016.pdf.
[23] Iyer K. IBM set to launch a 200-petaflop supercomputer by 2018[EB/OL]. [2016-09-22]. http://www.techworm.net/2016/06/ibm-set-launch-200-petaflop-supercomputer-2018.html.

Perspective on exascale-era computing

DENG Yuefan[1,2,3], ZHANG Lihao[1,2]

1. Stony Brook University, New York 11794, USA
2. Shandong Computer Science Center, National Supercomputer Centre in Jinan, Jinan 250101, China
3. School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China

Abstract Supercomputer has played its significant role in the development of science and technology. As we are moving forward into the exascale era, how to measure the performance of supercomputer is a crucial problem when concerning the system designs and programming models. Different benchmarks will give different conclusions. We introduce the three main ranking lists in supercomputing and their benchmarks, and analyze the current development and application perspective in different aspects.

Keywords supercomputer; Gorden Bell Prize; Top500; Green500; Graph500; TaihuLight; Tianhe-2

（责任编辑　刘志远）