ORNL/TM-2020/1561

An Analysis of System Balance and Architectural Trends Based on Top500 Supercomputers

Hyogi Sim
Awais Khan
Sudharshan S. Vazhkudai

August 11, 2020

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE US DEPARTMENT OF ENERGY

ORNL/TM-2020/1561

National Center for Computational Sciences

An Analysis of System Balance and Architectural Trends Based on Top500 Supercomputers

Hyogi Sim[†], Sudharshan S. Vazhkudai[†], Awais Khan[‡]

Oak Ridge National Laboratory[†], Sogang University[‡]

Date Published: August 2020

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831-6283
managed by
UT-Battelle, LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ABCI	AI Bridging Cloud Infrastructure
ACC	Accelerator
BB	Burst Buffer
BW	Bandwidth
CN	Compute Node
FLOPS	Floating Point Operations per second
HBM	High Bandwidth Memory (including G-DDR and HBM)
HDD	Hard Disk Drive
HPL	High Performance Linpack
HPC	High Performance Computing
GP-GPU	General Purpose Graphics Processing Unit
NFS	Network File System
NVM	Non-Volatile Memory
OLCF	Oak Ridge Leadership Computing Facility
ORNL	Oak Ridge National Laboratory
PFS	Parallel File System
SSD	Solid State Drive

ADDITIONAL FRONT MATERIAL

ACKNOWLEDGMENTS

We would like to thank Scott Atchley in the Technology Integration group for his valuable comments. The work was supported by, and used the resources of, the Oak Ridge Leadership Computing Facility, located in the National Center for Computational Sciences at ORNL, which is managed by UT Battelle, LLC for the U.S. DOE (under the contract No. DE-AC05-00OR22725).

ABSTRACT

Supercomputer design is a complex, multi-dimensional optimization process, wherein several subsystems need to be reconciled to meet a desired figure of merit performance for a portfolio of applications and a budget constraint. However, overall, the HPC community has been gravitating towards ever more FLOPS, at the expense of many other subsystems. To draw attention to overall system balance, in this paper, we analyze balance ratios and architectural trends in the world's most powerful supercomputers. Specifically, we have collected performance characteristics of systems between 1993 and 2018 based on the Top500 lists, and then analyzed their architectures from diverse system design perspectives. Notably, our analysis studies the performance balance of the machines, across a variety of subsystems such as compute, memory, I/O, interconnect, intra-node connectivity and power. Our analysis reveals that balance ratios of the various subsystems need to be considered carefully alongside the application workload portfolio to provision the subsystem capacity and bandwidth specifications, which can help achieve optimal performance.

1. INTRODUCTION

For several decades, supercomputers have provided the needed resources for modeling, simulation and data analysis in numerous scientific domains. The computing, storage and data resources offered by these systems have catered to both *capability*—requiring a large fraction of the machine—and *capacity*—needing medium-sized allocations—computing needs of applications [12]. The Top500 list [6] provides an excellent service to the HPC community by meticulously compiling the leading systems from the world based on the High Performance Linpack (HPL) benchmark [13], and publishing it bi-annually since 1993. The list reports key high-level architectural highlights (e.g., processor, interconnect type, memory, power, etc.) and FLOPS scores (R_{max} and R_{peak}).

Supercomputer design is a complex, multi-dimensional optimization process, in which several aforementioned vectors (and others such as storage) need to be reconciled in order to meet a desired *figure of merit* performance for a portfolio of applications and a budget constraint. For example, the goal of the Summit system at Oak Ridge National Lab (200 petaflop R_{peak}, 148.6 petaflop R_{max} and No. 1 in the June 2019 Top500 list) was to achieve a 5-10× performance improvement over its predecessor, Titan (the 27 petaflops system). In addition, the application workload mix has also been going through a transformation, with several supercomputing centers having to deal with new and emerging machine and deep learning codes, on top of the traditional modeling and simulation applications. Thus, during this process, it is natural that certain subsystems will be prioritized over certain others.

However, overall, the HPC community has been gravitating towards ever more FLOPS, at the expense of many other subsystems. While in theory it may seem obvious that a *balance* between the various subsystems is more important than just blindly prioritizing any one subsystem, in practice, however, this is seldom the case. Time and again, it is easier for centers to make a case for more FLOPS than for other subsystems. In reality, however, simply increasing the FLOPS may not improve application throughput if the other subsystems do not witness commensurate advances, as the end-to-end application performance is also dependent on other elements such as memory bandwidth, I/O throughput (for result and checkpoint data), and the like.

Therefore, what is needed is a careful consideration of the overall system balance and how the various subsystems reconcile with one another. System designers need to understand the trends not only within the individual subsystems but also with respect to one another. For example, one needs to understand the FLOPS trends in accelerator-based heterogeneous processors versus manycore processor architectures, but at the same time glean the nuances in FLOPS to memory bandwidth or memory capacity ratios; or memory bandwidth to intra-node connectivity bandwidth ratios; or file system to memory subsystem ratios; or interconnect to FLOPS ratios. Understanding the tradeoffs between the various subsystems will enable system designers to reconcile and provision them carefully, instead of producing suboptimal configurations that may be prone to performance bottlenecks.

In this paper, we conduct a detailed analysis of 26 years of Top500 lists since 1993, studying 10,708 supercomputers across several dimensions. Specifically, our contributions in this paper are as follows.

- We collect data from the Top500 lists and analyze detailed trends based on 10,708 supercomputers that have ranked in the list for the past 26 years between 1993 and 2018 (§ 4.1). We present performance and energy trends such as the following: the progressive increase in HPL scores over time, their comparison to Moore's law prediction, and the inflection point; the performance gap (factor) between the top systems and the lower-end systems; the historical trend in energy efficiency of systems and positions of

the No.1 systems; the trend in performance efficiency, i.e., the practical achievement of the theoretical peak performance by majority of the systems; and the commonly observed increasing trend in heterogeneous systems.

- We then select 27 systems, ranked in the top five in the past decade, i.e., between 2009 and 2018, and perform a deeper analysis on their architectural balance trends, including memory, file system, and interconnect (§ 4.2). We present the following results: the differences in the performance and energy efficiency of heterogeneous and traditional systems and the memory/core differences therein; the balance ratio between the memory subsystem and compute subsystem; the balance ratios between the memory, file system and the burst buffer subsystems; the balance ratios between network bisection and node injection bandwidth and the importance therein; and the correlation between interconnect performance and the over system performance efficiency.
- Lastly, we further select 15 heterogeneous machines from the 27 recent top five supercomputers and analyze the performance balance between the subsystem components for each recent heterogeneous system (§ 4.3). In this analysis, we particularly target the balance ratios and trends involved in newer technologies within a heterogeneous compute node such as multi-level memory and intra-node connectivity, both of which are essential in heterogeneous systems. We analyze the importance of memory (both DRAM and HBM) capacity and bandwidth per core and five different connections representing key intra-node links, and their relevance to different aspects of applications.

2. BACKGROUND: TOP500

In this section, we briefly introduce the Top500 project [6] and the resources it provides, which allow us to establish a basis for performing our analysis.

Since it was first launched in 1993, the Top500 project has been publishing a list of 500 of the world's most powerful supercomputers bi-annually, i.e., June and November in each year, on the project website [6]. Between 1993 and 2018, the project website has published 52 lists, which encompass 10,708 supercomputers from 2,894 institutions in the world. For compiling the list, the project evaluates supercomputers based on the High Performance Linpack benchmark (HPL) score, which assesses the runtime and accuracy of a distributed memory system in solving a dense linear system using double precision arithmetic [13]. Specifically, the participating supercomputers are ranked based on the number of floating point operations per second, or FLOPS. In addition to its semi-annual lists, the Top500 project also

Table 1. An example of the supercomputer specification from the Top500 data

Attribute	Example
Supercomputer	Summit
Installation site	DOE/SC/Oak Ridge National Laboratory
Total cores	2,397,824
Accelerator cores	2,196,480
Total memory capacity	2,801,664 GB
Processor type	IBM POWER9 22C 3.07GHz
Network interconnect family	Dual-rail Mellanox EDR Infiniband
Theoretical Peak (R_{peak})	200,795 TFlop/s
Linpack Performance (R_{max})	143,500 TFlop/s
Power consumption	9783 kW

publishes additional resources, e.g., useful statistics, interactive graphs, etc., via the project website. Particularly, the Top500 website publishes key specifications of individual supercomputers, e.g., processor type, memory capacity, interconnect family, etc., and such information, when combined with the semi-annual lists, can provide excellent insights on examining historical or recent trends in supercomputing [22, 16, 9, 8]. In this paper, we use the term *Top500 data* to refer to all available data that Top500 publicly publishes, including the semi-annual lists and the individual supercomputer specifications.

Table 1 shows an example specification of a supercomputer from the Top500 data. Particularly, the R_{peak} value is calculated based on the FLOPS values of all individual processing chips in the system, e.g., CPUs, GP-GPUs, etc., and demonstrates an ideal performance of the supercomputer without considering any potential overhead, e.g., network communication, data I/O, software algorithm, etc. In contrast, R_{max} is a measured score that has been acquired after running the HPL benchmark. Therefore, comparing the R_{peak} and R_{max} values provides a reasonable assessment of the overall processing efficiency of a supercomputer. For instance, the *Summit* supercomputer in Table 1, achieves approximately 71% of the ideal performance when running the HPL benchmark. Despite its abundance, the Top500 data lack comprehensive information about supercomputers, such as network bandwidth, file system performance, burst buffer capacity/performance, intra-node connectivity details, DRAM/HBM performance, etc., which is necessary for performing analysis on the architectural balance of a system. Therefore, we have collected extensive additional data through literature survey to fill in the gaps.

3. ANALYSIS OVERVIEW

In this section, we present our goals for analyzing the architectural trend of supercomputers based on the Top500 lists. Specifically, we perform analysis based on the following three analysis goals.

Overall performance trend (§ 4.1). Top500 adopts the High Performance Linpack (HPL) benchmark score [13] for normalizing performance and ranking supercomputers. However, the HPL score is a macro benchmark for measuring the aggregated processing power, and the score alone is a limited metric when it comes to unveiling the sophisticated architectural trends in supercomputers. We analyze the individual performance factors and find their correlations with the HPL scores.

Balance trends in recent supercomputers (§ 4.2). In this dimension, we perform a deeper analysis of the architectural trends and performance balance of the recent top five supercomputers on the Top500 list in the past decade. Specifically, we collect detailed information for each of the recent top supercomputers, and perform further analysis on the performance balance between the processing power and other subsystems in a supercomputer, e.g., memory, storage, burst buffer and network.

Balance trends in heterogeneous supercomputers (§ 4.3). Heterogeneous machines are becoming increasingly popular for achieving the desired system efficiency within the given budget and energy requirements [16]. We aim to identify key architectural trends and balance ratios from recent heterogeneous systems, e.g., intra-node connectivity and memory subsystem balance, and acquire insights for designing future systems.

For performing our analysis, we have collected available datasets from the Top500 website and also manually surveyed the detailed specification of individual target supercomputers for complementing the Top500 data.

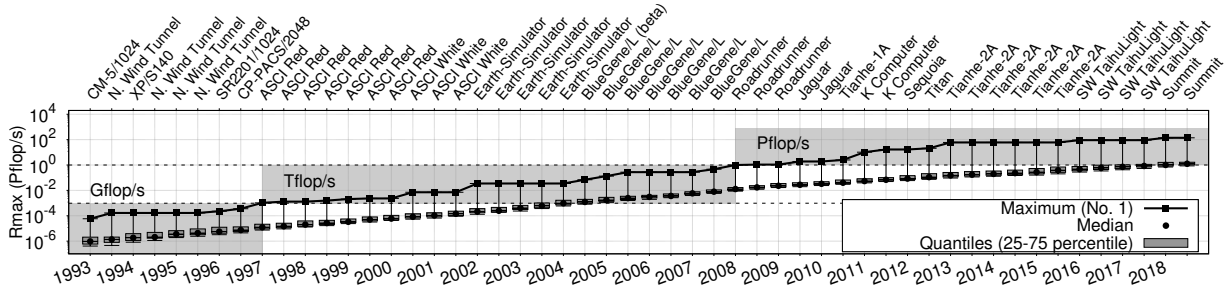


Figure 1. The growth of the linpack performance for the past 26 years (from 1993 to 2018). The graph depicts the HPL score distribution of 500 supercomputers for each year. The first tera-scale supercomputer (scored over 1 TFlop/s) was ASCI Red, and the first peta-scale supercomputer was Roadrunner.

4. ANALYSIS RESULTS

Based on the aforementioned goals (§ 3.), this section reports analysis results, namely, overall performance trend (§ 4.1), balance trends in recent supercomputers (§ 4.2), and performance balance in heterogeneous supercomputers (§ 4.3).

4.1 OVERALL PERFORMANCE TREND

We first study the overall performance trend in the Top500 list of systems over the past 26 years. Particularly, we analyze the trend in High Performance Linpack (HPL) scores of all 10,708 supercomputers that have appeared in Top500 between 1993 and 2018.

4.1.1 The growth of HPL scores

Figure 1 depicts the trend of R_{max} scores, i.e., the maximum observed performance (§ 2.), of all supercomputers that have appeared in the Top500 listings since 1993. We clearly observe a continuously increasing trend in performance over the past 26 years. On average, a newly introduced No.1 supercomputer has doubled the R_{max} score of its immediate predecessor. In addition, *ASCI Red* (1997) first recorded over a TFlop/s, while *Roadrunner* (2008) was the first petascale supercomputer. In Figure 2, we also compare the performance of No.1 machines against the prediction of Moore's Law [23]. Specifically, we normalize the R_{max} scores of No.1 machines based on the R_{max} score of the *CM-5/1024*, the No.1 machine in June 1993. We also project the ideal R_{max} scores based on the Moore's Law, i.e., the chip density and performance doubles every 18 months, using a dotted line. We observe that all No.1 machines since 1997 perform beyond the prediction of the Moore's Law. Particularly, the R_{max} score of *Tianhe-2A* in 2013

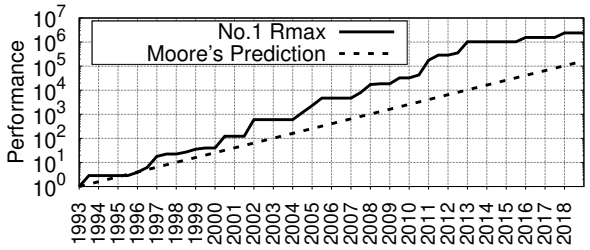


Figure 2. R_{max} of the No.1 supercomputers. The scores are normalized to the ideal projected score of Moore's Law [23] since 1993 and is shown by the dotted line.

exceeds the projection of Moore's Law by almost 100×. The most recent *Summit* supercomputer exhibits R_{max} that surpasses the projection by 18×. This demonstrates that the HPC systems address the physical limitation of the chip density by introducing multi-processing and heterogeneous architectures [24].

4.1.2 Low-end supercomputers

Another notable trend in Figure 1 is a highly skewed distribution of the R_{max} scores in all years, indicating a significant performance gap between high-end and low-end supercomputers. To articulate the trend, in Figure 3, we normalize R_{max} scores to the maximum score in each listing. We observe that 75% of the systems in each listing, i.e., 375 machines, scored at least an order of magnitude less than the No.1 supercomputer. The performance gap is widest in the June 2013 Top500 list, when the median HPL score of 500 systems was almost 400× lower than the score of *Tianhe-1A*. Although the performance gap is becoming narrower since then, the median HPL score in 2018 is still more than 100× lower than the top score.

4.1.3 Energy efficiency

One of the important metrics in evaluating system performance is energy efficiency, which is often measured by Flops per watt (W). Figure 4(a) shows the energy efficiency of clusters from the Top500 listings since 2005 [*]. We clearly observe an increasing trend in energy efficiency. Particularly, for each listing, the median energy efficiency of the corresponding 500 systems has increased by 1.2× on average. In addition, with the exception of 2005, the energy efficiency of the No.1 supercomputers is steadily positioned within the top 25%, demonstrating that the No.1 machines tend to run more energy efficiently than other machines. To further investigate this observation, we studied the correlation between the Top500 rank and energy efficiency, as shown in Figure 4(b). Each point in Figure 4(b) specifies the Pearson's correlation coefficient [†], where the energy efficiency is described as a function of the rank in the corresponding Top500 listing. We see that the strong negative correlation in earlier years, i.e., higher performance supercomputers being less energy efficient, is no longer the case in recent years (although no

[*]The earlier listings do not provide sufficient data about power consumption.

[†]The Pearson correlation coefficient, ρ, is defined as covariance of the variables (e.g., X and Y) divided by the product of their standard deviations, i.e., $\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$. A ρ value (ranging between -1 and 1) close to 0 indicates that no significant linear correlation is found.

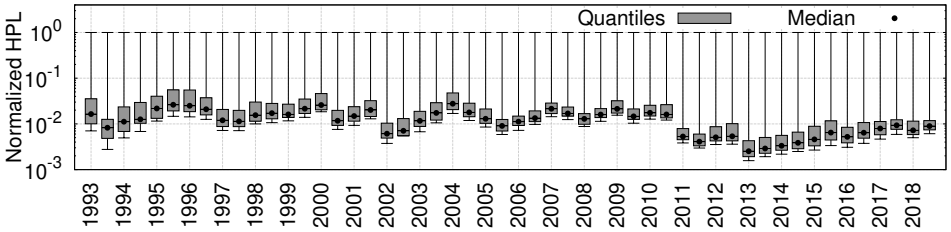
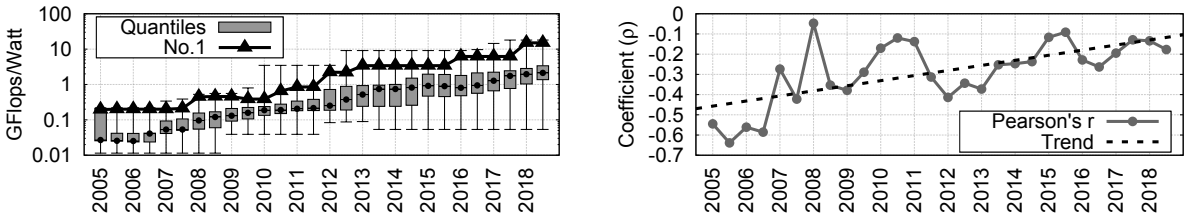


Figure 3. The distribution of normalized HPL scores in Top500. This clearly demonstrates a significant performance gap between the top and the rest supercomputers. In 2018, for instance, the HPL score of the No.1 supercomputer (*Summit*) is more than 100× greater than the median HPL score of the year.



(a) System power efficiency.

(b) Top500 rank and power efficiency.

Figure 4. Trends in the power efficiency in Top500 supercomputers. Over the past 26 years, the power efficiency of highly ranked supercomputers have been increasing.

positive correlation). Evidently, *Summit* (2018), the No.1 supercomputer in Top500, is also ranked No.2 in the Green500 [3] list for June 2019.

4.1.4 Performance efficiency

We now study the performance efficiency of systems, which we calculate as a ratio of R_{max} to R_{peak}, or $\frac{R_{max}}{R_{peak}}$ [16]. The average performance efficiency of 10,708 systems * is 0.67, indicating that most machines merely achieve less than 70% of their potential performance. Figure 5 further presents the annual trend in performance efficiency. In contrast to power efficiency (Figure 4), we do not observe an increasing trend in performance efficiency. Instead, on average, the median performance efficiency has decreased by about 4% each year. In addition, we also see that the performance efficiency of the No.1 supercomputers fluctuates heavily, which is a notable contrast to their power efficiency trend (Figure 4). For instance, the performance efficiency of the *K Computer* (2011) is 0.93, while 77% of No.1 supercomputers (40 out of 52) record performance efficiency scores below the overall median (0.67). Furthermore, performance efficiency in our analysis, which includes all systems in Top500, is about 15% lower than the earlier analysis with Top 10 supercomputers [16].

4.1.5 Achieving higher performance

A key factor in achieving a higher HPL score is to have a strong computing power. For this purpose, recent supercomputers tend to be equipped with a massive number of computing cores, as reported earlier in

* R_{peak} scores of some earlier supercomputers prior to 1994 are not available.

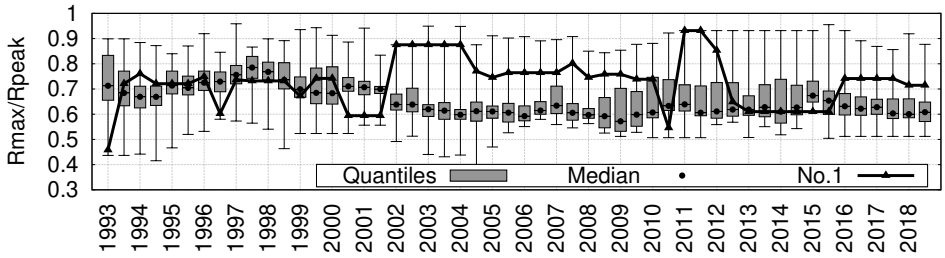


Figure 5. The trend of the performance efficiency, i.e., R_{max} : R_{peak}, in Top500 supercomputers. In contrast to the performance efficiency does not exhibit a clear increasing trend.

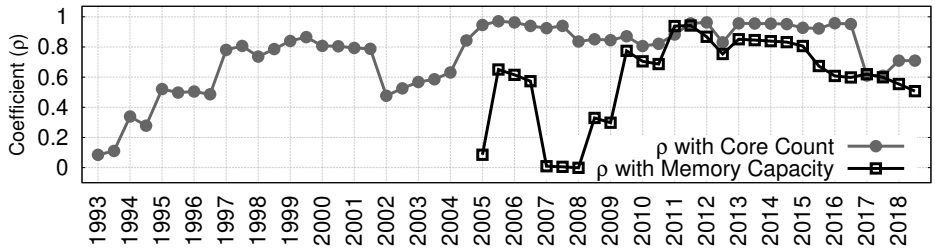


Figure 6. Trends in the correlation between performance and system attributes. Besides the number of cores, the memory capacity has also become a major factor to deliver a higher performance.

§ 4.1.1. Therefore, we now analyze how the total core count of a supercomputer affects its R_{max} score. Specifically, we performed a correlation analysis between R_{max} score and total core count for each year, as depicted in Figure 6. We observe the correlation coefficient (ρ) between HPL score and total core count is highest between 2013 and 2016, i.e., 0.95 on average. However, ρ drops drastically starting from 2017 that the average ρ between 2017 and 2018 is only 0.66, more than 30% lower than the previous year. One reason for this weaker correlation can be attributed to the increasing number of heterogeneous supercomputers, which we discuss further in § 4.3. In addition, Figure 6 also shows the correlation between HPL score and memory capacity. Starting from late 2009, the correlation between HPL score and memory capacity becomes noticeably higher, i.e., 0.74 on average between 2009 and 2018.

4.1.6 Heterogeneous supercomputers

Figure 7 shows the percentage of heterogeneous supercomputers, i.e., systems with additional accelerator processors such as GP-GPU, in the recent Top500 listings. For the past eight years, the number of heterogeneous systems in the listings has steadily increased, i.e., 1% or five systems annually, and they occupy about 28% (139 systems) in November 2018. We expect that this increasing trend will continue, particularly for addressing technological limitations (§ 4.1.1) and also for controlling the power consumption.

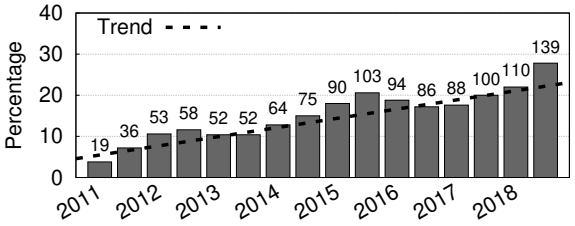


Figure 7. The increasing number of heterogeneous supercomputers in Top500 since 2011.

4.2 BALANCE TRENDS IN RECENT SUPERCOMPUTERS

In this section, we perform a deeper analysis on the performance trend in recent top supercomputers. Specifically, we focus on supercomputers that have ranked in the top five positions on the Top500 listings in the last decade, i.e., between 2009 and 2018. As summarized in Table 2, our target supercomputers consist of 15 heterogeneous (✦) and 12 traditional (❍) supercomputers.

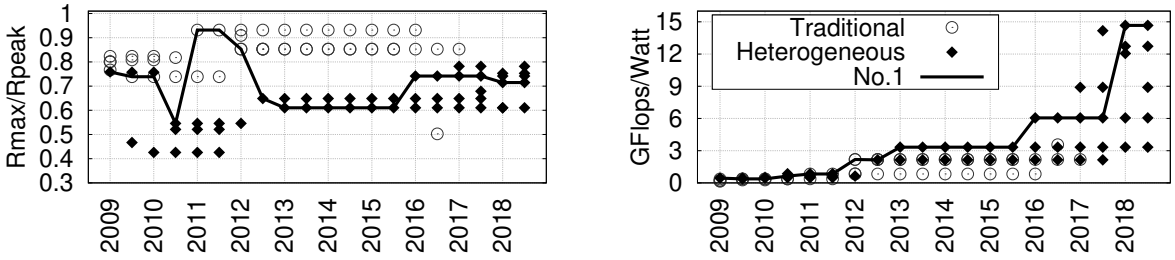
4.2.1 Overall system efficiency

Figures 8(a) and (b) show the performance efficiency (R_{max}:R_{peak}) and power efficiency (R_{max}:*Power*) of these supercomputers. We first observe that heterogeneous systems dominate the architectural trend in the top supercomputers. Particularly, since November 2017, all top five supercomputers are heterogeneous, indicating that the increasing popularity of the heterogeneous architecture (§ 4.1.6). Furthermore, in

Table 2. System characteristics of 27 supercomputers that have marked top five in Top 500 from 2009 to 2018.

	Top500 Rank																				Efficiency		Memory (M)		Storage				Network
	'09/06	'09/11	'10/06	'10/11	'11/06	'11/11	'12/06	'12/11	'13/06	'13/11	'14/06	'14/11	'15/06	'15/11	'16/06	'16/11	'17/06	'17/11	'18/06	'18/11	R_{max} to R_{peak}	R_{max} to Power	Cap. per Core	ΣBW to R_{max}	Cap (ΣM=1) PFS	Cap (ΣM=1) BB	BW (ΣM=1) PFS	BW (ΣM=1) BB	Bisection BW to ΣInjection BW
❍ **BlueGene/L**	5																				**0.80**	0.21	0.35	**1.18**	2.60	·	0.0000	·	0.0038
✦ **Roadrunner.1**		2	3																		**0.76**	0.44	5.98	0.27	28.56	·	0.0001	·	0.0627
✦ **Roadrunner.2**	❶																				**0.76**	0.45	5.98	0.27	26.97	·	0.0001	·	0.0593
❍ **Jaguar.1**	2																				**0.77**	0.15	2.05	0.34	34.13	·	0.0005	·	0.0072
❍ **Pleiades**	4																				**0.80**	0.23	1.00	0.05	**139.26**	·	**0.0017**	·	·
❍ **JUGENE**	3	4	5																		**0.82**	0.36	0.50	**0.98**	14.22	·	0.0000	·	0.0046
✦ **Jaguar.2**		❶	❶	2	3	3															**0.74**	0.38	1.07	0.25	32.80	·	0.0004	·	0.0142
❍ **Kraken**		3	4																		**0.81**	0.27	1.52	0.23	22.99	·	0.0001	·	·
✦ **Tianhe-1**		5																			0.47	0.37	1.55	**0.79**	9.46	·	0.0003	·	·
✦ **Nebulae**			2	3	4	4															0.43	0.49	2.22	0.41	2.49	·	0.0001	·	·
✦ **Tsubame-2.0**				4	5	5															**0.52**	0.85	1.34	0.59	59.90	1.72	0.0001	0.0005	**1.2291**
✦ **Tianhe-1A**				❶	2	2	5														**0.55**	0.64	2.92	0.25	8.36	·	0.0003	·	·
❍ **Hopper**				5																	**0.82**	0.36	1.45	0.41	9.44	·	0.0001	·	·
❍ **K Computer**					❶	❶	2	3	4	4	4	4	4	5							**0.93**	0.83	2.00	0.46	22.31	**8.18**	0.0001	0.0002	0.0741
❍ **Sequoia**							❶	2	3	3	3	3	3	4	4	5					**0.85**	2.18	1.00	0.20	36.67	·	0.0004	·	0.1221
❍ **Mira**							3	4	5	5	5	5	5								**0.85**	2.18	1.00	0.20	46.67	·	0.0001	·	0.0682
❍ **Super MUC**							4														**0.91**	0.85	2.00	0.29	53.33	·	0.0003	·	0.2778
❍ **JUQUEEN**								5													**0.85**	2.18	1.00	0.20	0.22	·	0.0001	·	0.0112
✦ **Titan**							❶	2	2	2	2	2	3	3	4	5					**0.65**	2.14	2.37	0.20	44.30	·	0.0002	·	**1.1158**
✦ **Tianhe-2A**								❶	❶	❶	❶	❶	❶	2	2	2	2	4	4		**0.61**	3.32	8.00	0.07	5.83	·	0.0002	·	0.1918
✦ **SW TaihuLight**															❶	❶	❶	❶	2	3	**0.74**	6.05	16.00	0.09	8.00	·	0.0000	·	0.1094
❍ **Cori**																5					**0.50**	3.56	1.66	0.19	27.40	1.83	0.0001	0.0003	0.4814
✦ **Piz Daint**																	3	3		5	**0.78**	**8.91**	2.23	0.06	46.06	·	0.0001	·	**0.7703**
✦ **Gyoukou**																		4			**0.68**	**14.17**	**33.94**	0.03	24.67	·	**0.0013**	·	·
✦ **ABCI**																			5		**0.61**	**12.06**	12.82	0.13	41.32	3.26	0.0004	**0.0008**	**0.6995**
✦ **Summit**																			❶	❶	**0.71**	**14.67**	9.64	0.13	**88.59**	2.62	0.0001	0.0004	**1.0222**
✦ **Sierra**																			3	2	**0.75**	**12.72**	7.52	0.14	**110.00**	**5.06**	0.0001	**0.0005**	0.5120

(a) ❍ and ✦ indicate that the corresponding supercomputer has homogeneous or heterogeneous architectures, respectively.
(b) The color intensity shows the comparison between values within the corresponding column.

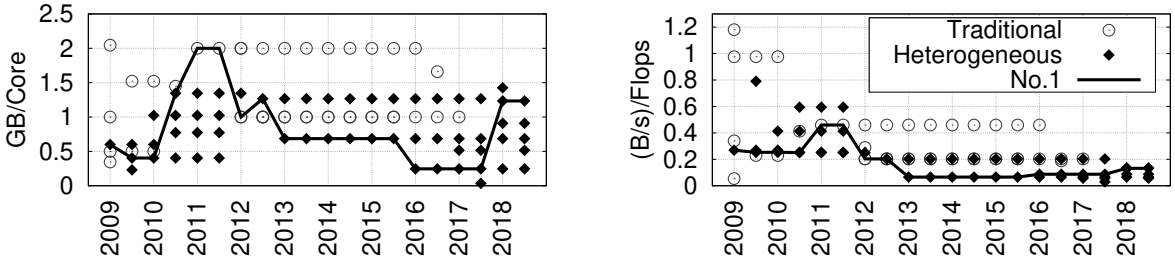


(a) Performance efficiency.

(b) Power efficiency.

Figure 8. Trends of performance and power efficiency in recent top five supercomputers. The heterogeneous architecture clearly improve the power efficiency but also imposes challenges to increase the performance efficiency.

Figure 8(a), we notice that heterogeneous systems tend to exhibit a lower performance efficiency, i.e., achieving less than 80% of the theoretical peak performance (R_{peak}). In contrast, Figure 8(b) shows that the power efficiency of heterogeneous systems far exceed that of traditional systems, especially since 2017. Specifically, the average power efficiency of the heterogeneous machines (5.5 GFlops/Watt) is about five times higher than the average power efficiency of the traditional machines (1.1 GFlops/Watt). Our observation clearly demonstrates the benefit, i.e., energy efficiency, and also challenges, i.e., technical obstacles to realize the potential performance [11], of the heterogeneous architecture.



(a) Memory capacity per core.

(b) Memory bandwidth per Flops.

Figure 9. Performance balance in system memory. Despite the increasing performance of the memory system, the per Flops memory bandwidth has decreased due to the growth of the processing power.

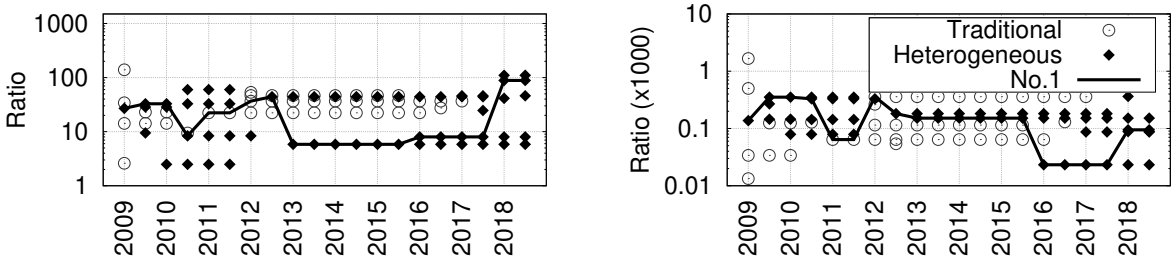
4.2.2 System memory

Next, we analyze the performance trend in the memory subsystem. For heterogeneous systems, the memory capacity and bandwidth are the sums of the DRAM and HBM capacity and bandwidth. First, Figure 9(a) shows the trend in the memory capacity per core ($\Sigma Memory_{Cap}$:$\Sigma Cores_{Total}$) of recent top machines. We observe that most systems are clustered around 1 GB in the graph. Only three supercomputers, i.e., *Jaguar.1*, *K Computer*, and *Super MUC*, furnish more than 2 GB of memory per processing core. In addition, the per-core memory capacity of heterogeneous supercomputers (0.7 GB on average) tend to be lower than the per-core memory capacity of traditional systems (1.3 GB on average), although the heterogeneous systems tend to be equipped with a greater amount of system memory (more than 300 TB on average). This indicates that the increase in the core count from accelerators, e.g., GP-GPU, is greater than the increase of memory (HBM) from accelerators in the heterogeneous machines. In fact, in the heterogeneous systems, the average HBM capacity per accelerator core is merely 0.2 GB, about 14× less than the average DRAM capacity per CPU core (3.3 GB).

Next, Figure 9(b) depicts the performance balance between the aggregate memory bandwidth and the peak processing power ($\Sigma Memory_{BW}$:R_{peak}) of the target supercomputers. Overall, we clearly see a diminishing trend in the balance ratio, indicating that the processing power grows faster than the system memory speed. For instance, the highest ratio value in 2019, i.e., 0.13 from *Summit*, is about 9× lower than the highest ratio in 2009, i.e., 1.2 from *BlueGene/L*. Further, after 2011, none of the top systems exceed 0.5 B/s per Flops (more on this in § 4.3). This observation conforms to the limitation of provisioning the memory bandwidth in the modern processor design [15].

4.2.3 Parallel File System

Most supercomputers are equipped with a networked parallel file system (PFS) to support capacity requirements of running applications. The main memory is inevitably used as a buffer space for manipulating datasets in the PFS. Therefore, we analyze the performance balance between the PFS and the memory subsystem. Figure 10(a) and (b) show the capacity and bandwidth ratios between PFS and memory subsystem, i.e., PFS_{Cap}:$\Sigma Memory_{Cap}$ and PFS_{BW}:$\Sigma Memory_{BW}$, respectively. Note that we only consider scratch file systems that parallel applications primarily exploit for storing data, i.e., excluding NFS */home* and archival storage areas. For the file system capacity (Figure 10(a)), we observe that the ratio values are scattered between 2 and 100, except for two systems, i.e., *Pleiades* and *Gyoukou*, which provide substantially larger file system space compared to their memory capacity, i.e., 140× and 410×, respectively.



(a) Capacity ratio.

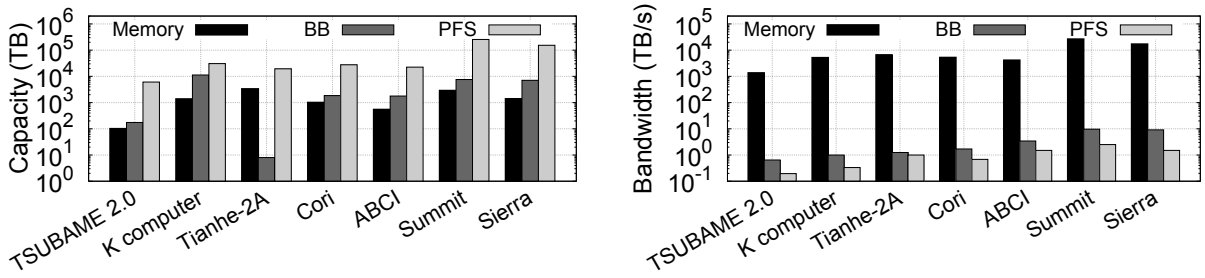
(b) Bandwidth ratio.

Figure 10. Performance balance between file system and memory subsystem. We do not observe a drastic change in the file system capacity and bandwidth. On average, the file system capacity and bandwidth are about 44× larger and 13,353× slower, respectively, than the system memory capacity in the recent top five supercomputers.

The overall average ratio is 44, meaning that the recent top supercomputers tend to provision the PFS capacity to be 44× larger than their memory capacity. *Summit* has a ratio of 89, almost 2× greater than the overall average. Note that a smaller capacity ratio between the PFS and memory requires a more frequent purge operations to guarantee a sufficient capacity in the PFS, while a larger capacity allows a longer retention of data in the PFS. Similar to the capacity ratio, we do not observe a clear change over time in the bandwidth ratio (Figure 10(b)). On average, the file system bandwidth in the recent top systems are 13,353× lower than the aggregated memory bandwidth, although we have observed significant variance (σ=17,000) among these systems. The PFS in *Summit* is about 10,000× slower than its aggregated memory speed, justifying a burst buffer.

4.2.4 Burst Buffer Storage

The burst buffer (BB) has recently become popular to mitigate the performance gap between memory and file system [19]. Seven out of the 27 recent top systems (Table 2) have BB storage, either within a compute node or in a dedicated set of nodes, e.g., IO forwarding nodes, inside the cluster. In Figure 11, we compare the (a) capacity and (b) bandwidth of the aggregated system memory, BB, and PFS of each of these seven systems, i.e., (a) $\Sigma Memory_{Cap}$:PFS_{Cap}:BB_{Cap} and (b) $\Sigma Memory_{BW}$:PFS_{BW}:BB_{BW}, respectively. From Figure 11(a), we see that the BB capacity of most machines range between the capacity of memory and



(a) Storage capacity.

(b) Storage bandwidth.

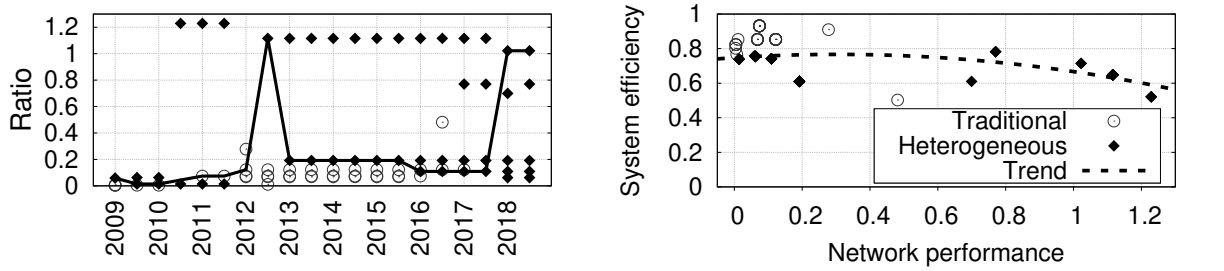
Figure 11. Burst buffer characteristics in seven recent supercomputers. In these supercomputers, the burst buffer capacity is about 3× larger than the system memory, and its bandwidth is about 3× faster than the bandwidth of the parallel file system.

PFS except for *Tianhe-2A*, which employs SSDs in its 256 IO forwarding nodes [28]. On average, the BB capacity is about 3× larger than the memory capacity, and the *K Computer* exhibits the highest ratio, i.e., 8× larger than the memory capacity. Similarly, the BB bandwidth also ranges between the memory bandwidth and the PFS bandwidth, as depicted in Figure 11(b). However, the bandwidth gap between memory and BB is noticeably large in all seven systems. On average, the BB bandwidth in the seven systems is about 3.2× greater than the PFS bandwidth but also about 3,065× slower than the total memory bandwidth. In addition, compared to the earlier systems (e.g., *K Computer*, *Cori*, etc.), *Summit* and *Sierra* provide a significantly higher BB bandwidth (i.e., 9.7 TB/s and 9.1 TB/s respectively) with a less number of compute nodes and SSDs.

BBs are much lower in capacity compared to the PFS and can typically accommodate 2-3 snapshots of a system memory checkpoint (e.g., *Summit's* 512GB of DRAM compared to 1.6TB of node-local SSD.) Another emerging provisioning strategy is to combine the salient properties of a BB (high rates) and a PFS (better reliability and capacity) into a single flash-based storage tier (e.g., the *Perlmutter* system at NERSC in 2020). While it can offer better rates, a high-capacity, all-flash tier will be cost prohibitive (Perlmutter's all-flash PFS offers 4TB/s but only around 30PB). The intent is for such a tier to be backed by a project or a campaign storage with larger capacity. On the flip side, future systems such as OLCF's *Frontier* system in 2021 will continue to provide a node-local flash-based BB and an HDD-based PFS, with 2-4x capacity and bandwidth compared to OLCF's *Summit* BB and PFS, respectively (BB: 7.4PB, 9.7TB/s; PFS: 250PB, 2.5TB/s; the PFS also caters to medium-term analysis needs like a project store). Consequently, the deep-storage hierarchy on the high-end systems is still evolving to better fit the various usage scenarios at the respective centers.

4.2.5 Interconnect network

The interconnect performance is a crucial factor that affects the capability of a supercomputer when it comes to processing large-scale, inter-node jobs. We summarize the networking performance characteristics of the 27 recent top supercomputers in Figure 12. Note that we could not find the bisection bandwidth information from seven systems (marked 'NA' in Table 2) and exclude such systems in Figure 12. First, Figure 12(a) shows the ratio between the bisection bandwidth and the total injection bandwidth ($NetworkBW_{Bisection}$: $\Sigma NetworkBW_{Injection}$), demonstrating how efficiently the global interconnection network of a supercomputer can handle the communication requests from individual



(a) Bisection bandwidth to total injection bandwidth.

(b) Network performance and HPL performance efficiency.

Figure 12. Performance trend in the interconnect network. (a) shows the interconnect network performance in processing all-to-all communication. (b) demonstrates that the interconnect network performance does not exhibit a strong correlation to the HPL performance efficiency.

compute nodes at the full scale. We observe that the bisection bandwidth in most systems are substantially lower than the total injection bandwidth, i.e., the aggregated injection bandwidth from all compute nodes. On average, the bisection bandwidth is 32% of the total injection bandwidth for the 20 systems. However, three supercomputers, i.e., *Tsubame-2.0* (ratio of 1.2, non-blocking fat tree), *Titan* (1.1, 3D torus), and *Summit* (1.0, non-blocking fat tree), show bisection bandwidth exceeding the total injection bandwidth, indicating that the bisection bandwidth in these systems does not impose a bottleneck in global communications such as all-to-all communication. Although it is ideal to design a system bisection bandwidth to suffice the total injection bandwidth, but it needs to be weighed against design factors, e.g., target application communication profile, budget, etc.

Next, Figure 12(b) shows the correlation between this interconnect performance, i.e., the ratio of the bisection bandwidth to the total injection bandwidth, and the overall performance efficiency, i.e., $R_{max} : R_{peak}$ (§ 4.1.4). We do not find any strong correlation between the overall performance efficiency and the interconnect network performance. This weak correlation suggests that the network performance does not substantially impact the ability to acquire a high score in the HPL benchmark. However, depending on the target environment and mission, attaining a high bisection bandwidth for a system may be necessary. For instance, a recent analysis of the five-year job log from *Titan* suggests that over 54% of the CPU hours were consumed by large-scale jobs (using more than 2,048 compute nodes) even though 90% of the submitted jobs were using less than 256 compute nodes [27]. In such an environment, a sufficient bisection bandwidth is essential for supporting large-scale jobs.

4.3 PERFORMANCE BALANCE IN HETEROGENEOUS SUPERCOMPUTERS

In this section, we analyze the performance balance in intra-node connectivity of the 15 heterogeneous supercomputers from the 27 top recent supercomputers (§ 4.2). For each heterogeneous supercomputer, we

Table 3. Performance balance ratio in the 15 recent heterogeneous supercomputers.

	CN Flops		Intranode Connectivity				System Efficiency	
	CPU (GFlops)	ACC (GFlops)	CPU-CPU (GB/s)	CPU-ACC (GB/s)	ACC-ACC (GB/s)	RSD (σ:μ)	Performance (R_{max}:R_{peak})	Power (GFlops/Watt)
R.Runner.1	14.4	435.2	12.80	2.00	25.60	**1.75**	**0.76**	**0.44**
R.Runner.2	14.4	435.2	12.80	2.00	25.60	**1.75**	**0.76**	**0.45**
Jaguar.2	288.4	665.0	·	8.00	·	**3.17**	**0.74**	**0.38**
Tianhe-1	270.0	224.0	11.20	8.00	8.00	**3.08**	**0.47**	**0.37**
Nebulae	127.6	515.2	12.80	8.00	8.00	**0.87**	**0.43**	**0.49**
Tsubame-2.0	152.0	1,545.0	12.80	8.00	8.00	**1.72**	**0.52**	**0.85**
Tianhe-1A	140.6	515.0	12.80	8.00	8.00	**5.39**	**0.55**	**0.64**
Titan	144.2	1,341.4	·	8.00	·	**2.28**	**0.65**	**2.14**
Tianhe-2A	422.4	5,033.2	16.00	15.75	·	**2.15**	**0.61**	**3.32**
SW TaihuLight	95.0	3,040.3	16.00	·	16.00	**1.75**	**0.74**	**6.05**
PizDaint	166.4	4,812.8	·	15.75	·	**2.22**	**0.78**	**8.91**
Gyoukou	332.8	23,511.0	·	15.75	15.75	**0.46**	**0.68**	**14.17**
ABCI	3,840.0	28,672.0	20.80	15.75	50.00	**5.52**	**0.61**	**12.06**
Summit	1,105.9	43,008.0	64.00	50.00	50.00	**2.06**	**0.71**	**14.67**
Sierra	1,105.9	28,672.0	64.00	75.00	75.00	**2.04**	**0.75**	**12.72**

[a] CN Flops column shows the breakdown of the Flops performance between CPUs and accelerators (ACC) in a compute node.
[b] RSD column lists the relative standard deviation from bandwidth of main memory, HBM, CPU-to-CPU, CPU-to-ACC, ACC-to-ACC, and network injection.
[c] A smaller RSD value indicates a smaller bandwidth variance among those intra-node connections.

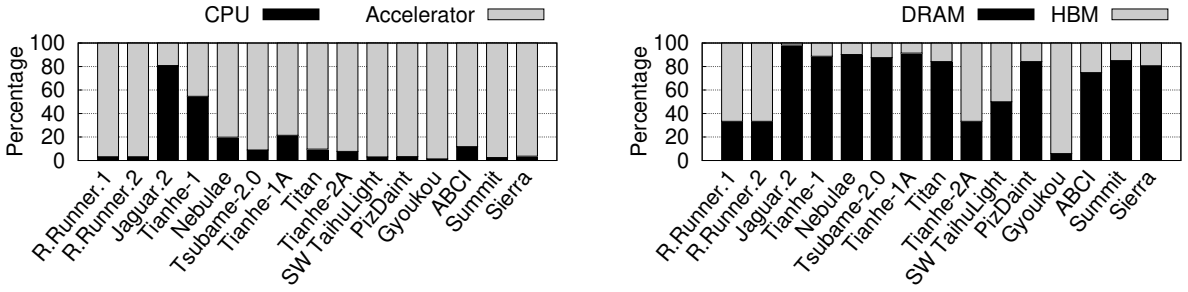
further summarize important characteristics of the intra-node connectivity in Table 3.

4.3.1 Provisioning Accelerators

We first analyze the proportion of accelerators in the overall system performance for the 15 heterogeneous supercomputers. Figure 13(a) depicts the Flops (R_{peak}) ratio between the conventional CPU and the accelerators for each heterogeneous system ($\Sigma Flops_{CPU}$:$\Sigma Flops_{ACC}$). It is clearly noticeable that the accelerator dominates the overall performance in most heterogeneous systems. For the 15 heterogeneous systems, the accelerators contribute to 84% of the system R_{peak} on average, and *Jaguar.2* is the only machine wherein the accelerators produce less than 50% of the system R_{peak}. However, *Jaguar.2* was in a partial upgrade phase from Cray XT5 to XK6 in November 2009 (Table 2) and thus only 960 out of 18,688 compute nodes had GP-GPUs [10]. Recent *Summit* and *Sierra* systems rely on the accelerator for more than 95% of overall system Flops. This indicates that it is essential to utilize the accelerators efficiently to fully exploit the processing power of heterogeneous supercomputers.

Figure 13(b) shows the capacity between DRAM (for CPUs) and HBM (for accelerators), i.e., $\Sigma DRAM_{Cap}$:ΣHBM_{Cap}. Despite the strong dominance of the accelerators in R_{peak}, the DRAM capacity still dominates the HBM capacity in many heterogeneous systems. On average, DRAM provides 68% of total system memory capacity. Besides the higher cost of HBM, this is also because the CPUs require more memory for arbitrating the tasks among accelerators and also for handling other system demands, e.g., running the operating system. In contrast, most accelerators primarily perform computational tasks. In addition, systems may also be provisioning more DRAM to accommodate CPU-only jobs. For instance, even on heterogeneous systems, there is a significant fraction of CPU-only jobs due to slower adoption of GPUs (e.g., GPU adoption on the Titan supercomputer was only 28% in 2018 [27]) or some codes may not be amenable to the GPU and the system may need to support them anyway. While such jobs will not be using the full potential of the system, it may be necessary for the system to accommodate them in its portfolio. In such cases, one approach to still effectively utilize the node would be to multiplex CPU-only jobs and GPU-based jobs. For example, one can co-locate the post-processing analysis of an end-to-end job (simulation + data analysis) on the same CPU/GPU node, wherein a GPU-based simulation is multiplexed with the CPU-based analysis in an in-situ fashion [18].

In Figure 13(b), only four heterogeneous systems, i.e., *Roadrunner.1*, *Roadrunner.2*, *Tianhe-2A*, and

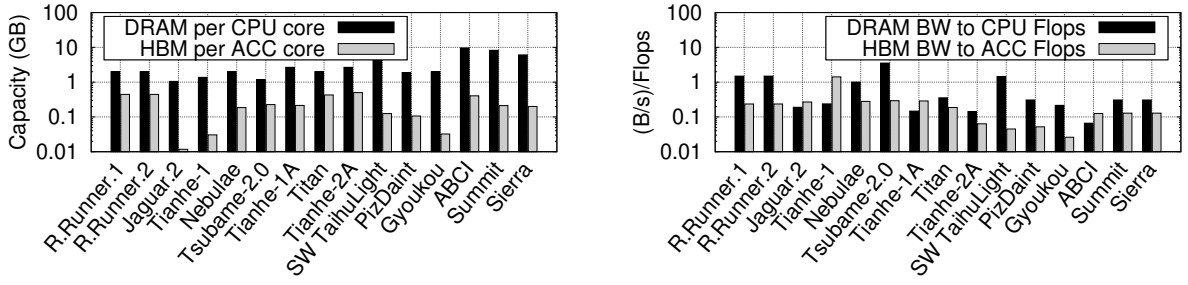


(a) Flops ratio.

(b) Memory ratio.

Figure 13. Provisioning the accelerators. (a) shows the ratio of the system Flops (R_{peak}) between CPUs and accelerators. (b) shows the capacity ratio between system main memory and HBM.

Gyoukou, feature more amount of HBM than the amount of DRAM. Interestingly, these four machines are equipped with accelerators that are not GP-GPUs. For instance, *Gyoukou* is equipped with the PEZY-SC2 accelerators [5], and the accelerator memory provides 95% of the overall memory capacity. Similarly, *Roadrunner* and *Tianhe-2A* adopt the IBM PowerXCell 8i processor and the in-house developed Matrix2000, respectively, for their accelerators.



(a) Memory capacity.

(b) Memory bandwidth.

Figure 14. The performance balance of memory subsystem in 15 recent heterogeneous supercomputers. The per-core memory capacity if about 15× higher for CPUs. Also, despite the higher memory bandwidth of HBMs, the bandwidth to Flops ratio is lower for accelerators due to their higher Flops count.

4.3.2 Memory Subsystem

In § 4.2.2, we have studied the performance trend in system memory for 27 recent top supercomputers. In a heterogeneous architecture, however, accelerators are commonly installed with a dedicated memory system that can be independent to the system main memory. Therefore, for the 15 heterogeneous supercomputers, we separately analyze the performance balance of the two different memory types, i.e., the system main memory for CPUs and the HBM for accelerators. First, Figure 14(a) shows the main memory capacity per CPU core ($\Sigma DRAM_{Cap}$:$\Sigma Cores_{CPU}$) and the HBM capacity per accelerator core (ΣHBM_{Cap}:$\Sigma Cores_{ACC}$) for the 15 heterogeneous supercomputers. Noticeably, the per-CPU core memory capacity (3.5 GB on average) is significantly larger, i.e., about 15×, than the per-accelerator core memory capacity (0.2 GB on average). In addition, the per-CPU core memory capacity is particularly large in *Sunway TaihuLight* (8 GB), *ABCI* (9.6 GB), *Summit* (11.6 GB), and *Sierra* (5.8 GB). As mentioned earlier in § 4.3.1, this dissimilarity in the per-core memory capacity is attributed to the fundamental difference between CPUs and accelerators in the processing architecture and target tasks. Further, HBM is also more expensive than DRAM, which will likely limit its capacity.

To address such cost constraints, future systems may also consider deeper memory hierarchies, wherein HBM and DRAM is supplemented with NVM (e.g., more HBM and very little to no DRAM, but with a large node-local, byte-addressable NVM like 3D XPoint). Technologies are becoming available that can directly populate GPU's HBM from the node-local SSDs using GPUDirect methods, obviating the need to load data onto DRAM and then copy to the GPU memory. However, this needs to be weighed against the need to accommodate CPU-only jobs that will need enough DRAM. In any case, memory hierarchies are likely to get even richer. While applications prefer a flatter, easily addressable memory address space, budget constraints will eventually influence how deep and wide the memory hierarchy gets.

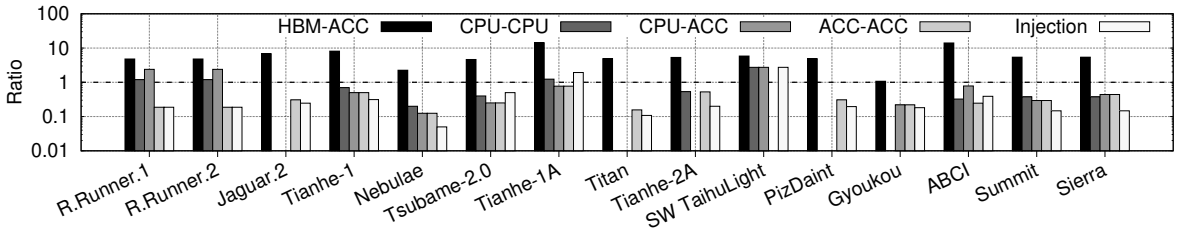


Figure 15. Balance of the intra-node connectivity in 15 recent heterogeneous supercomputers. The graph shows the bandwidth of each internal connection normalized to the system main memory bandwidth. ACC denotes an accelerator such as GP-GPU.

Figure 14(b) depicts the memory bandwidth per Flops for CPUs and accelerators ($\Sigma DRAM_{BW}$:$\Sigma Flops_{CPU}$ and ΣHBM_{BW}:$\Sigma Flops_{ACC}$). Here, we calculate the ratio of aggregated HBM bandwidth to the aggregated Flops of accelerators (Table 3). Except for four supercomputers, i.e., *Jaguar.2*, *Tianhe-1*, *Tianhe-1A*, and *ABCI*, the DRAM bandwidth to CPU Flops is about 3× greater than the HBM bandwidth to accelerator Flops. However, this does not indicate the DRAM bandwidth is generally higher than the HBM bandwidth, but is because of the higher processing power of accelerators (Flops count), as specified in Table 2 and 3.

4.3.3 Intra-node connectivity

In a heterogeneous supercomputer, a compute node houses additional hardware, e.g., GP-GPU, HBM, which requires additional connections, e.g., data exchange between CPU and GP-GPU (denoted as ACC), inside the node. Such internal connections, or intra-node connectivity, should be designed carefully to prevent performance bottlenecks within a compute node. Therefore, we analyze the balance in the intra-node connectivity for 15 heterogeneous systems. Figure 15 shows the bandwidth of five internal connections namely HBM-to-ACC bandwidth, CPU-CPU bandwidth, CPU-ACC bandwidth, ACC-ACC (peer-to-peer) bandwidth and injection bandwidth. All bandwidth values are normalized to the system main memory bandwidth of the corresponding supercomputer. A missing bar indicates that the corresponding connection is not applicable to the system. For instance, each compute node in *Titan* has a single CPU and GPU, and thus CPU-to-CPU and ACC-to-ACC connections do not exist. However, each node in Summit has two IBM P9 CPUs with CPU-CPU connectivity via IBM's X-Bus, CPU to DRAM connectivity, six Nvidia Volta GPUs with HBM, resulting in HBM-to-ACC and ACC-ACC connectivity (NVLink), and CPU-ACC (NVLink) links. Overall, most internal connections within a compute node are slower than the system main memory bandwidth, except for the HBM-to-ACC and the ACC-to-ACC bandwidth. On average, the HBM-to-ACC bandwidth is 6.2× greater than the main memory bandwidth, while the ACC-to-ACC bandwidth is almost comparable (i.e., 0.9×) to the main memory bandwidth). In addition, the average CPU-to-CPU, CPU-to-ACC, and network injection bandwidth are 0.8×, 0.3×, and 0.5×, respectively, of the main memory bandwidth. Since the HBM-to-ACC bandwidth is 6.2x DRAM bandwidth, it might appear that the DRAM bandwidth is the bottleneck in transferring data between the CPU and the ACC; however, it should be noted that the CPU-to-ACC (e.g., PCIe or NVLink) bandwidth is 0.3× DRAM bandwidth, indicating that it is in fact the slower link in the end-to-end data path.

An important measure for assessing the balance of the intra-node connectivity is the variance among the multiple connections. In Table 3, the RSD column lists the relative standard deviation * of main memory,

*For a standard deviation (σ) and a mean (μ), the relative standard deviation (RSD) is $\frac{\sigma}{\mu}$.

CPU-to-CPU, CPU-to-ACC, ACC-to-ACC, and network interconnect bandwidth. According to the RSD values (lower means better balance), *Nebulae* (RSD=0.87) and *Gyoukou* (RSD=0.46) exhibit a well-balanced intra-node connectivity. In contrast, *Tianhe-1A* (RSD=5.39) and *ABCI* (RSD=5.52) show the most skewed intra-node connectivity ratios among the 15 heterogeneous supercomputers. For the 15 heterogeneous supercomputers, the ACC-to-ACC connection exhibits the largest impact on the performance efficiency of the HPL benchmark, i.e., $\frac{R_{max}}{R_{peak}}$, compared to the other individual connections. Specifically, the correlation coefficient (ρ) between the ACC-to-ACC bandwidth and the performance efficiency is about 0.6, about 2× greater than the average from all internal connection bandwidth values, i.e., the average ρ from the main memory (ρ=0.1), HBM (ρ=0.3), CPU-to-CPU (ρ=0.4), CPU-to-ACC (ρ=0.3), ACC-to-ACC (ρ=0.6), and network injection bandwidth (ρ=0.1). This is because the HPL benchmark is a compute-intensive task [13], for which accelerators, e.g., GP-GPUs, are heavily utilized in heterogeneous supercomputers (§ 4.3.1). Likewise, the HBM bandwidth (ρ=0.3) affects more than the main memory bandwidth (ρ=0.1) does for HPL. Recent technologies, such as NVLink [14] and Infinity Fabric [17, 7], directly address this observation, i.e., the necessity for fast communication among CPUs and accelerators, by introducing a fast and specialized interconnect for accelerators instead of relying on the generic PCIe interconnect.

It is more important to provision for the eventual application workload than to simply achieve a balance across all of the intra-node connections. While a low RSD implies better balance across the links, it is more important to better provision the links that will get utilized more, even it results in a higher RSD. Of course, care should be taken to not let any one connection lag behind too much. Therefore, provisioning of intra-node connectivity should carefully consider the application portfolio, their demands on the CPU/ACC and the associated memory, the anticipated data movement between the CPU and ACC and between the ACCs, and the potential cost to efficiently specify the bandwidth. For example, if the workload is expected to transfer more data between the processors, it will be more important to provision a higher CPU-ACC bandwidth compared to the other links, etc.

5. RELATED WORK

With the past 26 years of semi-annual reporting, the TOP500 [6] project has become the most reliable, up-to-date source for studying the leading technical trends of the world's most powerful supercomputers. Particularly, Top500 adopts the High Performance Linpack (HPL) benchmark [13] to normalize and rank the performance of supercomputers. Due to its long history and abundant resources, several prior reports have studied historical and architectural trends in supercomputing by analyzing the data from the Top500 project. For instance, an earlier report in 2001 [22] summarized the supercomputing history based on the Top500 data. A study in 2008 [21] also provided statistical summaries of supercomputer architectures and future performance predictions based on the Top500 data. Similarly, a recent study [16] analyzed the architectural trend of supercomputers until 2012, and anticipated the future trends based on the past tendency. Compared to such prior studies, this paper not only provides the most up-to-date analysis of its kind but also performs a deeper analysis for revealing the trend in the performance balance, which is often overlooked in prior reports.

There are other ranked lists for complementing the sole performance metric of HPL [25], including the Gordon Bell Prize [1] (focused on application performance), IO500 [4] (specialized in the I/O performance), Green500 [3] (assessing the power efficiency), and Graph500 [2] (measuring the parallel graph processing capability). Despite their usefulness, we do not include such projects in this study

especially due to insufficient resources and history compared to the Top500 project.

There exist a few studies that have addressed the increasing architectural complexity in supercomputers and the consequent importance of the performance balance in the system design [20, 26]. For instance, an earlier study [20] indicated that the performance of subsystem components in a supercomputer, e.g., memory, disk, network, etc., should be comparable to the processing performance of CPU. However, the study is dated and thus does not consider recent technologies such as accelerators or burst buffers. A recent study [26] analyzes the architecture and the performance balance in three Department of Energy (DOE) supercomputers, i.e., Titan, Summit, and Sierra. Despite its technical details, the study only discusses the architectures of the three aforementioned supercomputers and is limited for demonstrating the overall trend in supercomputing. Similarly, there exist other studies [11, 19] that primarily analyzed a single performance aspect of supercomputers, e.g., accelerator, file system, interconnect network, etc. In contrast, this paper thoroughly analyzes the architectural trend and performance balance in memory subsystem, file system, interconnect network, and intra-node connectivity in recent supercomputers.

6. CONCLUSION

In this paper, we have analyzed over 10,000 supercomputers from Top500, and presented recent architectural trends in leading supercomputers. Furthermore, we have analyzed the performance balance trends for the top supercomputers in the past decade. Particularly, our analysis is focused on revealing the trend in the performance balance, which has been disregarded in the prior analysis reports. We believe that our analysis will provide a useful guideline to understand the architectural trends in leading supercomputers and also to design next generation supercomputers.

7. REFERENCES

References

[1] *ACM Gordon Bell Prize*. https://awards.acm.org/bell.

[2] *Graph 500 | large-scale benchmarks*. https://graph500.org/.

[3] *Green500 | TOP500 Supercomputer Sites*. https://www.top500.org/green500/.

[4] *IO-500 [Virtual Institute for I/O]*. https://www.vi4io.org/std/io500/start.

[5] *ZettaScaler - WikiChip*. https://en.wikichip.org/wiki/zettascaler.

[6] *TOP500 Lists*. http://www.top500.org/lists/.

[7] *Radeon Vega 20 Will Have XGMI*. https://www.phoronix.com/scan.php?page=news_item&px=AMDGPU-XGMI-Vega20-Patches.

[8] S. R. Alam, J. A. Kuehn, R. F. Barrett, J. M. Larkin, M. R. Fahey, R. Sankaran, and P. H. Worley. Cray XT4: An Early Evaluation for Petascale Scientific Simulation. In *SC '07: Proceedings of the 2007 ACM/IEEE Conference on Supercomputing*, 2007.

[9] B. Austin, C. Daley, D. Doerfler, J. Deslippe, B. Cook, B. Friesen, T. Kurth, C. Yang, and N. J. Wright. A metric for evaluating supercomputer performance in the era of extreme heterogeneity. In *2018 IEEE/ACM Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS)*, Nov 2018.

[10] Arthur S Bland, Jack C Wells, Otis E Messer, Oscar R Hernandez, and James H Rogers. Titan: Early Experience with the Cray XK6 at Oak Ridge National Laboratory. In *Proceedings of cray user group conference (CUG 2012)*. Cray User Group Stuttgart, Germany, 2012.

[11] Neal E Davis, Robert W Robey, Charles R Ferenbaugh, David Nicholaeff, and Dennis P Trujillo. Paradigmatic Shifts for Exascale Supercomputing. *The Journal of Supercomputing*, 62(2), 2012.

[12] Douglas W Doerfler, Mahesh Rajan, Marcus Epperson, Courtenay T Vaughan, Kevin Pedretti, Richard Frederick Barrett, and Brian Barrett. A Comparison of the Performance Characteristics of Capability and Capacity Class HPC Systems. Technical report, Sandia National Lab., Albuquerque, NM, 2011.

[13] Jack J Dongarra, Piotr Luszczek, and Antoine Petitet. The LINPACK Benchmark: Past, Present and Future. *Concurrency and Computation: practice and experience*, 15(9), 2003.

[14] D. Foley and J. Danskin. Ultra-Performance Pascal GPU and NVLink Interconnect. *IEEE Micro*, 37 (2), Mar 2017.

[15] A. Kagi, J. R. Goodman, and D. Burger. Memory Bandwidth Limitations of Future Microprocessors. In *23rd Annual International Symposium on Computer Architecture*, ISCA '96, 1996.

[16] P. M. Kogge and T. J. Dysart. Using the TOP500 to Trace and Project Technology and Architecture Trends. In *SC '11: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '11, 2011.

[17] Kevin Lepak, Gerry Talbot, Sean White, Noah Beck, Sam Naffziger, et al. The next generation amd enterprise server product architecture. *IEEE hot chips*, 29, 2017.

[18] M. Li, S. S. Vazhkudai, A. R. Butt, F. Meng, X. Ma, Y. Kim, C. Engelmann, and G. Shipman. Functional Partitioning to Optimize End-to-End Performance on Many-core Architectures. In *SC '10: Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis*, Nov 2010.

[19] Ning Liu, Jason Cope, Philip Carns, Christopher Carothers, Robert Ross, Gary Grider, Adam Crume, and Carlos Maltzahn. On the role of burst buffers in leadership-class storage systems. In *IEEE 28th Symposium on Mass Storage Systems and Technologies (MSST)*, 2012.

[20] John McCalpin. Memory bandwidth and machine balance in high performance computers. *IEEE Technical Committee on Computer Architecture Newsletter*, 12 1995.

[21] Hans Werner Meuer. The Top500 Project. Looking Back over 15 Years of Supercomputing Experience. *PIK-Praxis der Informationsverarbeitung und Kommunikation*, 31(2), 2008.

[22] Yoshio Oyanagi. Future of Supercomputing. *Journal of Computational and Applied Mathematics*, 149(1), 2002.

[23] R. R. Schaller. Moore's law: past, present and future. *IEEE Spectrum*, 34(6), June 1997.

[24] Amar Shan. Heterogeneous Processing: a Strategy for Augmenting Moore's Law. *Linux Journal*, 2006(142), 2006.

[25] Erich Strohmaier, Hans W Meuer, Jack Dongarra, and Horst D Simon. The Top500 List and Progress in High-Performance Computing. *Computer*, 48(11), 2015.

[26] Sudharshan S. Vazhkudai and Bronis R. et. al. de Supinski. The Design, Deployment, and Evaluation of the CORAL Pre-exascale Systems. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*, SC '18, 2018.

[27] F. Wang, S. Oral, S. Sen, and N. Imam. Learning from Five-year Resource-Utilization Data of Titan System. In *Workshop on Monitoring and Analysis for High Performance Computing Systems Plus Applications*, HPCMASPA '19, 2019.

[28] Weixia Xu, Yutong Lu, Qiong Li, Enqiang Zhou, Zhenlong Song, Yong Dong, Wei Zhang, Dengping Wei, Xiaoming Zhang, Haitao Chen, Jianying Xing, and Yuan Yuan. Hybrid Hierarchy Storage System in MilkyWay-2 Supercomputer. *Front. Comput. Sci.*, 8(3), June 2014.