On-Going Study: Analysis of the Characteristics and Development Trends of the Next-Generation of Supercomputers

September 6, 2017

Earl Joseph, ejoseph@idc.com
Steve Conway sconway@idc.com
Bob Sorensen, bsorensen@idc.com

Project Requirements

1. Gather information on pre-exascale and exascale systems today and through 2028
2. Concentrate on major HPC developer countries: US, China, EU, Japan, others?
3. Build database of technical information on the research and development efforts on these next-generation machines
4. Collect information on the flow of funding (amount from the country to the companies, etc.)

Data Point Collected on Each System

- Comparisons of System Attributes
  - System Attributes: Planned Performance
  - System Attributes: Architecture and Node Design
  - System Attributes: Power
  - System Attributes: MTBF Rates
  - System Attributes: KPIs
- Comparisons of Prices
  - Comparison of System Prices
  - Who's Paying for It?
- Comparisons of Ease-Of-Use
  - Ease-of-Use: Planned New Features
  - Ease-of-Use: Porting/Running of New Codes on a New Computer
  - Ease-of-Use: Missing Items that Reduce Ease-of-Use
  - Ease-of-Use: Overall Ability to Run Leadership Class Problems

Data Points (Continued)

- Comparisons of Hardware Attributes
  - Hardware Attributes: Processors
  - Hardware Attributes: Memory Systems
  - Hardware Attributes: Interconnects
  - Hardware Attributes: Storage
  - Hardware Attributes: Cooling
  - Hardware Attributes: Special Hardware
  - Hardware Attributes: Estimated Utilization
- Comparisons of Software Attributes
  - Software Attributes: OS and Special Software
  - Software Attributes: File Systems
  - Software Attributes: Compilers and Middleware
  - Software Attributes: Other Software
- Comparisons of Supporting Research & Development
  - R&D Plans
  - R&D Plans: Partnerships

The Computers Evaluated in The Study (As of Today)

HPC Systems Under Study

Historical Reference

HPC Exascale Chronology

Computer Name	Planned Delivery Date/ Estimated
Hazel Hen	2015
SuperMUC	2015
CORI	2016, 1Q
TaihuLight	2 Q 2016
Piz Diant	3Q 2016
Cheyenne	2017, 1Q
Summit (OLCF4)	Install 2017, 3Q,
Taihu Prototype	2017 4Q
TianHe2 A	2017, 4Q
Sugon	
NSF	2Q, 2018
Sierra (ATS-2)	2018, 3Q
UK Three System Upgrade	2018, 3Q
Tianhe-3	3Q 2018
Aurora	2018, 4Q

Computer Name	Planned Delivery Date/ Estimated
Falke (Falcon)	2019
NERSC-9	2020, 4Q
Crossroads (ATS-3)	2020, 4Q
Exascale System	2020, 3Q
Sunway 2020	2020.4Q
NUDT 2020	prototype in 2018, full version in 2020, 4Q
OLCF5	2022
ATS-4	2023
NERSC-10	2024
ATS-5	2025
NERSC-11	2028
Source: Hyperion 2017	

HPCs In the Study (To Date)

HPCS Pre-Exascale and Exascale

Computer Name	Overview	Prime Developer /Industry Partner	Location	Organization	Country
Hazel Hen	Cray XC40-system	Cray	University of Stuttgart	HLRS	Germany
SuperMUC	2015: Phase 2, 2011-2013 3 Phase 1 upgrades	Lenovo/IBM	Leibniz Supercomputing Centre, Munich	LRZ	Germany
CORI	NERSC-8	Cray	Wang Hall	NERSC, LBNL	USA
TaihuLight	Sunway	NRCPC	National Supercomputing Center in Wuxi	Jiangsu Province, the city of Wuxi and Tsinghua University,	China
Piz Diant		Cray	Lugano, Switzerland	Swiss National Supercomputing Centre CSCS	Switzerland
Cheyenne		SGI		NCAR	USA
Summit (OLCF4)	CORAL	IBM	(OLCF)	ORNL	USA

HPCs In the Study (To Date)

HPCS Pre-Exascale and Exascale

Computer Name	Overview	Prime Developer /Industry Partner	Location	Organization	Country
Taihu Prototype	Sunway Exascale Prototype	(NRCPC)	National Supercomputing Center in Jinan		China
TianHe2 A	TianHe2 upgrade	NUDT/Inspur		NUDT	China
Sugon	Exascale Prototype	Sugon /AMD		Dawning Information Industry, Chinese Academy of Sciences	China
NSF	Leadership-Class Computing Facility - Phase 1 (Phase 2 will be exascale)		Pittsburg Supercomputer Facility (?)	NSF-sponsored	USA
Sierra (ATS-2)	CORAL	IBM, NVIDIA, Mellanox	LLNL	LLNL	USA
UK Three System Upgrade	S&T Facilities Council	IBM		University of Edinburgh, ECMWF, Daresbury Lab	UK
Tianhe-3	NUDT Exascale Prototype	NUDT	National Super Computer Tianjin Center		China
Aurora	CORAL	Intel/Cray	ANL	ANL	USA
Falke (Falcon)	Hazel Hen Follow-on	NA	High Performance Computing Center, University of Stuttgart	HLRS	Germany
NERSC-9	APEX 2020	TBD	Wang Hall LBNL	NERSC,	USA

HPCs In the Study (To Date)

HPCS Pre-Exascale and Exascale

Computer Name	Overview	Prime Developer /Industry Partner	Location	Organization	Country
Crossroads (ATS-3)	APEX 2020	TDB	Strategic Computing Complex (SCC), Building 2327	LANL	USA
Exascale System	Tera 1000 Follow-on	CEA/ Bull		CEA/ Bull	France
Sunway 2020	Sunway 2020	NRCPC		Sunway 2020	China
NUDT 2020	TianHe-3 Prototype Follow-on			NUDT 2020	China
OLCF5	Summit Follow-On		Oak Ridge National Lab	ORNL	USA
ATS-4				LLNL	US
NERSC-10				NERSC	US
ATS-5				LANL	US
NERSC-11				NERSC	US

Snap Shot of the Database

Summary Attributes						
Computer Name	**Summit (OLCF4)**	**Taihu Prototype**	**TianHe2 A**	**Sugon**	**NSF**	**Sierra (ATS-2)**
Overview	CORAL	Sunway Exascale Prototype	TianHe2 upgrade	Exascale Prototype	Leadership-Class Computing Facility - Phase 1 (Phase 2 will be exascale)	CORAL
Prime Developer /Industry Partner	IBM	National Research Center of Parallel Computer Engineering and Technology (NRCPC) and the ,	NUDT/Inspur	Sugon /AMD		IBM, NVIDIA, Mellanox
Location	Oak Ridge Leadership Computing Facility (OLCF)	National Supercomputing Center in Jinan			Pittsburg Supercompuer Facility (?)	LLNL
Organization	ORNL		NUDT	Dawning Information Industry, Chinese Academy of Sciences	NSF-sponsored	Lawrense Livermore National Lab
Country	USA	China	China	China	USA	USA
Summary System Attributes						
Planned Delivery Date/ Estimated	Install 2017, 3Q,	2017 4Q	2017, 4Q		2Q, 2018	2018, 3Q
Planned Performance PF/ Estimated PF*	200		200-300 (100+ LP)	2.5 PF		120-150
GF/Watt Goal	11.5 to 13.2		5-Apr			11.5 - 13.2
System Base Design	Custom		~ 18,000 nodes, Dual-socket FT2000/64 server equipped with two Matrix2000 GPDSP (6 teraflops per node)			Custom
Breadth of Applicability				Wide range of applications such as HPC, deep learning, big data and cloud computing.	At least 80 percent of the capacity of the Phase 1 system will be allocated to scientists and engineers through NSF's Petascale Computing Resource Allocation (PRAC) program.	
Market Impact						
Node Configuration	>4600 nodes, 40 TFLOPS/node		Phytium FT-2000/64 ARM chip (512 GFLOPS) + Matrix 2000 GPDSP	hyper-converged self-adaptive parallel system architecture, local processor based high-performance compute nodes		>1500 fat nodes, 40 TFLOPS/node
Memory/Node	512 GB DDR4 +HBM, 800 GB of NVRAM		U/U/C			512 GB DDR4 +HBM, 800 GB of NVRAM

An Overview of the Findings

There Are A Number Of Major Leadership-class HPCs In Development

Development is under way across a wide range of major HPC suppliers and regions including China, the EU, Japan, and the United States.

- Most of these systems are pre-exascale designs: systems that will underwrite much of the technology critical to the development of the hardware and software necessary to support the spate of exascale systems planned for the 2020 to 2022 time-frame.
- As such, the bulk of the systems planned for the next four years target a peak performance capability between 10 and 300 teraflops, with the bulk of the lower-end systems closer to completion this year or the next, while higher performance systems are targeted for completion closer to 2020.

Many Architectures Under Development

There is a wide range of different architectural design paths to an exascale system.

- Some projects are looking to partner with a commercial vendor, such as Cray or IBM, to help them develop a leadership-class system in keeping with the overall product offerings of their commercial partner.
- Others, such as NUDT's Tianhe-2 A group in China, are essentially looking to custom-build a system that likely will be produced in very limited quantiles, be used primarily in domestic markets, and developed with little expectation of eventual commercialization.
- In addition, it is clear that there is no agreed upon architectural scheme for these pre-exascale systems.

Power Consumption is a Major Concern

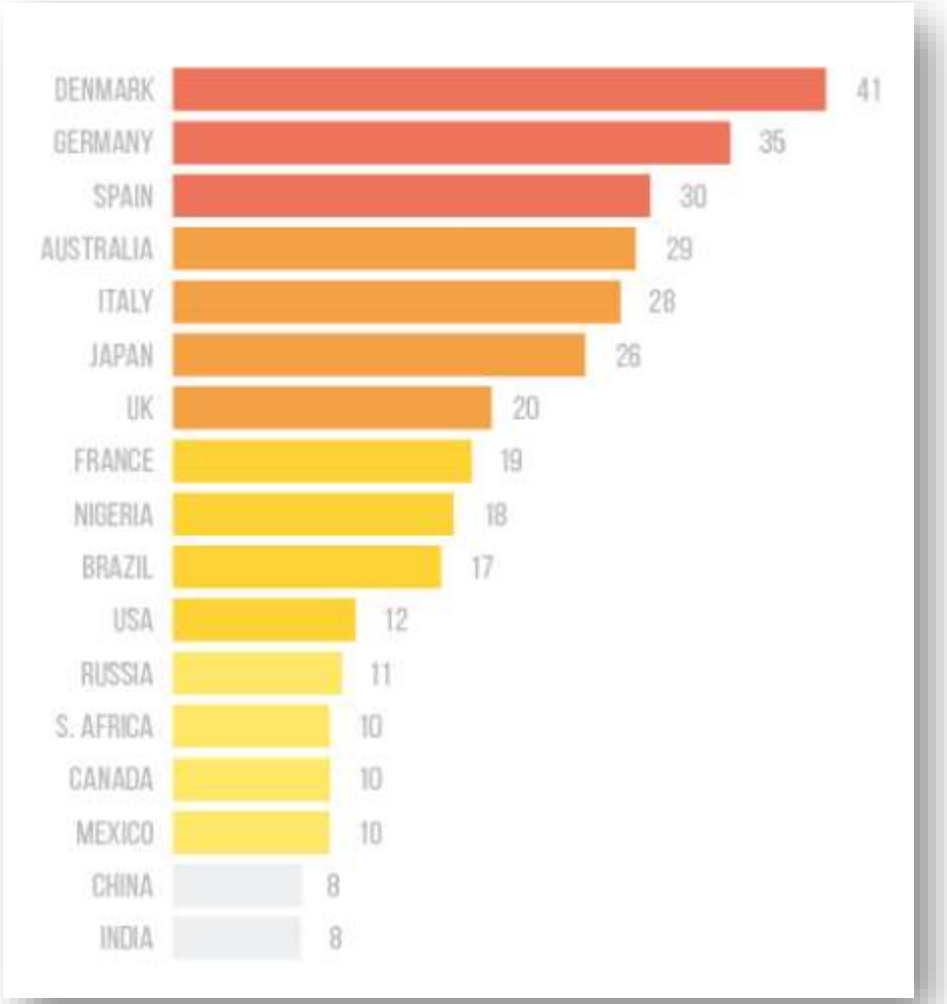
Concerns about power consumption and related efficiencies are keeping total power needs for pre-exascale –and exascale - systems generally below 30 MW.

Hyperion is currently gathering power and cooling data and insights from some of the most important HPC centers in the US and overseas.

Summary So Far

- Respondents generally expect to pay about $1 million/megawatt/year for a system that will be around 25MW.
  - But this can vary significantly by location---see next chart
  - All are seeking PUE's of <1.1
- One time installation costs vary, as some sites had existing capabilities and were only looking at upgrades.
  - One estimated a one time complete new total power and cooling cost of $150-$250 million
  - Another said a power and cooling upgrades could cost ~$50 million

Average National Energy Costs (in US cents/kWh)



If China = India = 1

Then

{

US = 1.5

France = 2.75

UK = 2.5

Japan = 3.25

Germany = 4.374

}

Most Aren't Focused on Peak Performance Anymore (Maybe…)

The designers, developers, and users of pre-exascale systems are not generally concerned with the theoretical peak computation performance of their new systems.

- There is an increased emphasis on determining the ability of a new system to deliver a sustained performance that captures the ability of a system's overall compute, memory, interconnect, and storage infrastructure to execute an end-to end task - over one that stresses pure computationally capability.

There Is A Wide Mix in Exascale Budgets

Pre-exascale designers are operating under a wide range of budgets from a low of $25 million to well over ten times that amount.

- Some of the most expensive pre-exascale systems – such as the most technologically, aggressive one-off systems - are projected to cost $250 million or much more. These systems represent some of the most advanced HPC developments in the world, and include significant non-recoverable engineering (NRE) costs.
- Others, primarily those that are one step behind the leading-edge of performance, generally are looking at budgets an order of magnitude less.
  - Many of these systems have less aggressive NRE requirements and instead rely primarily on hardware and software technology supplied by their vendor partners.

Lots of New Hardware Trends

- There is a wide range of pre-exascale processors and related GPU accelerators being considered for inclusion in the various systems. Deep Learning will accelerate this trend.
- The overarching trend in pre-exascale design is toward more memory, more SSDs, and the use of additional memory accelerators, such as burst buffers or high bandwidth memory packages as a way to deal with the increasing need for higher bandwidth and lower latency memory systems.
- For most of the systems that will be delivered soon, designers are opting for either InfiniBand, Intel OmniPath, or in a few cases, a custom in-house interconnect scheme.
- Leadership class supercomputer designs have overall storage requirements that are moving well into the 100PB range in the next few years.

Some New Software Trends, Some New Game Changers

- Linux, in its many variants, has become the stock operating system for most leadership-class supercomputer, and Hyperion analysts assess that this will be the case for at least the next five years.
- Lustre and GPFS are and likely will continue to be the major file system software for leadership-class supercomputers for at least the next five years.
- There is increasing attention being paid to non-traditional HPC software that Hyperion Research analysts expect will become increasingly important in the next few years, such as big data infrastructures, virtualization schemes such as Docker, and deep learning.

Access the First Full Exascale Report Competed Last Year

http://www.aics.riken.jp/en/overview/report

HYPERION RESEARCH

QUESTIONS?

ejoseph@hyperionres.com

sconway@hyperionres.com

bsorensen@hyperionres.com

mthorp@hyperionres.com