Harvesting Tomorrow's Technology for Today's Problems

Matthew T. Ziegler

Director – HPC Architecture and Strategy

Lenovo Data Center Product Group

2016 Lenovo Internal. All rights reserved.

Executive Overview

HPC MISSION STATEMENT

Use Lenovo's Innovation Engine and Technical Expertise to help solve our customer's most complex problems by leveraging Lenovo's broad portfolio of systems, deep skills, leadership scale of economics and open partnerships to deliver the benefits of Open HPC to all customers.

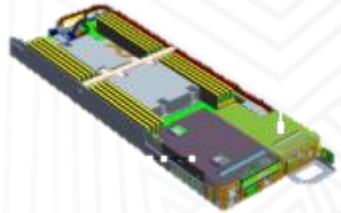
- Recently achieved 92 total entries on Top500 = #2 vendor
- Broad range of Lenovo systems types listed
- Over 178k servers on the list
- Largest OmniPath system listed

- Large systems continue to use homogenous designs
- Small/Med. systems are heterogenous/pay-as-you-grow
- Supercomputers are pushing away from x86
- Transactional systems leverage current Xeon technology

- Accelerators becoming rapidly adopted for high end
- Bootable accelerators changing landscape
- Dense GPGPUs emerging in both ML/DL and HPC

- OEMs are adopting new technology regardless of revenue forecast
- Increased pressure to carry latest IT tech/solutions
- ROI is difficult to understand outside of x86

Lenovo System x Dense – A History of Platform Changes



2016+ Dense Optimized

- As Dense market matures, Lenovo intends to continue platform innovation.
- Targets General Purpose and HPC with 2U4N design
- Provides flexible node, rack and datacenter designs

NeXtScale 2013

- Designed to continue innovation into the dense server market started by iDataPlex
- Adopts more open standards approach to compute, storage and networking
- First to be Ubuntu certified
- Standard Rack installable
- Lowest platform cost to date

2009 iDataPlex

- Launched to capture the web 2.0 datacenter emerging market
- First system to explore high density compute design and energy efficiency
- Deployed with Intel board design with no HPC features
- 2U Chassis with no exotic midplane

BladeCenter 2008

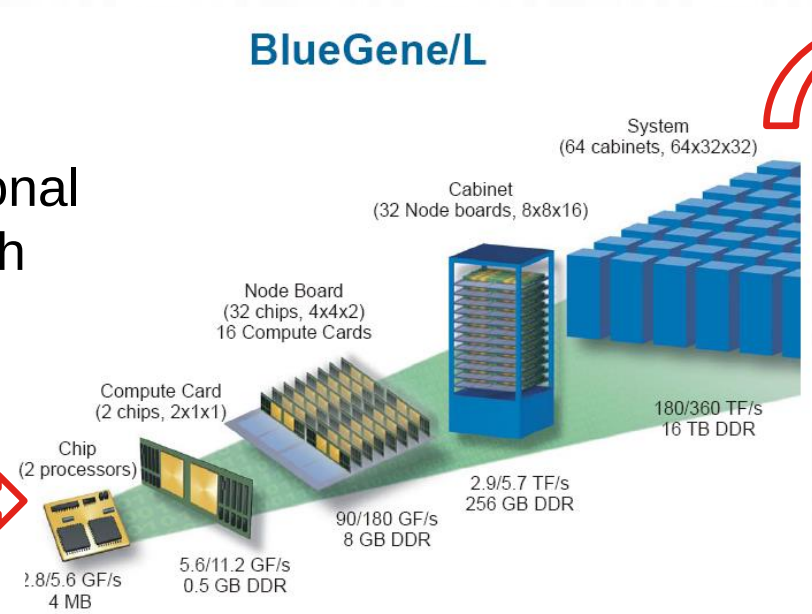
- First to a PetaFlop!
- PetaFlop was first benchmark only with BladeCenter H AMD Opteron systems and DDR IB
- Built from off-the-shelf bladecenter
- Second phase employed Cell technology
- Birthplace of xCAT 2.0

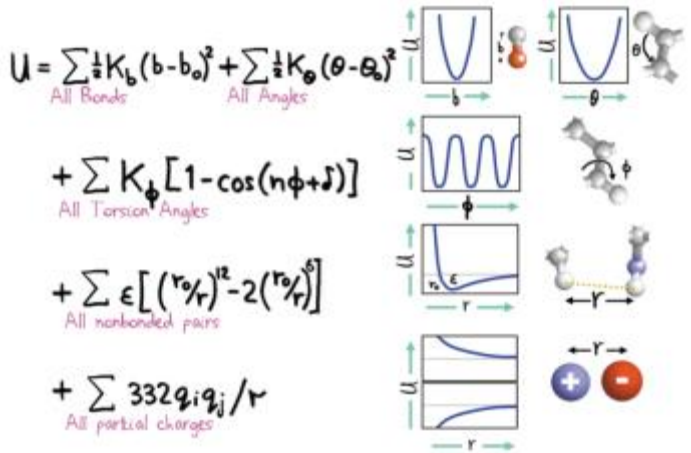
2000 Netfinity

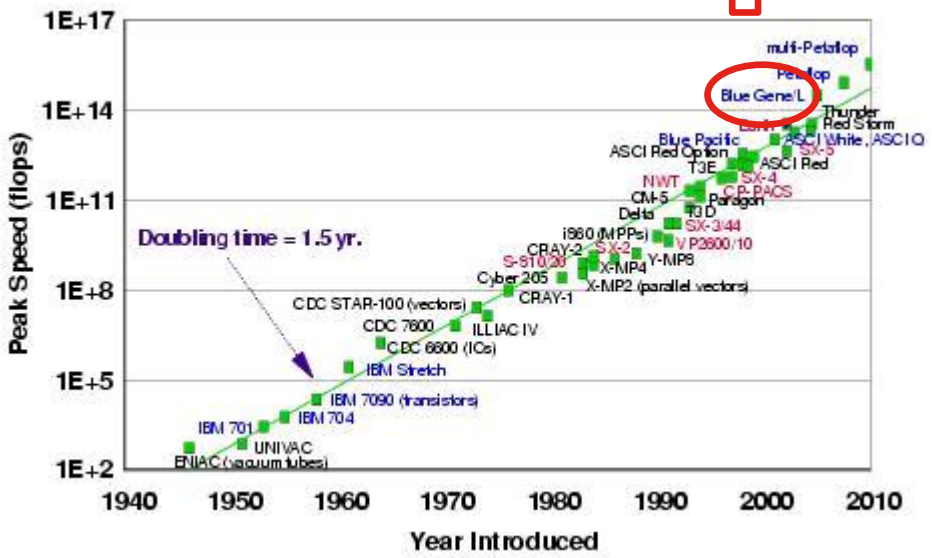
- System x's first x86-based Linux Cluster landed at #80 on the Nov. 2000 Top500 list.
- Built from 256 General Purpose Netfinity Servers with a GbE backbone

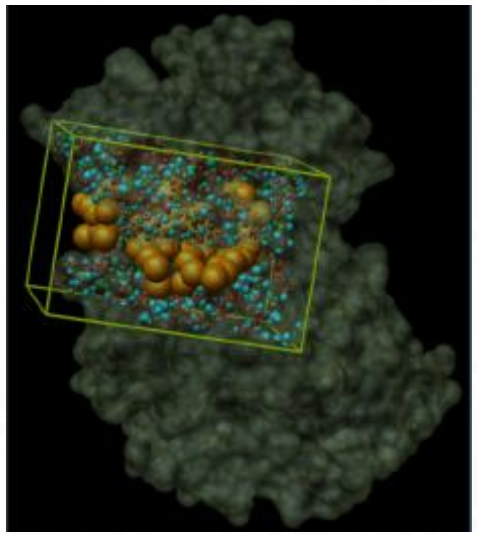
IT Innovation was driven by Life Sciences

The BlueGene architecture was designed to solve the computational intensive problem associated with protein folding









2016 Lenovo Internal. All rights reserved.

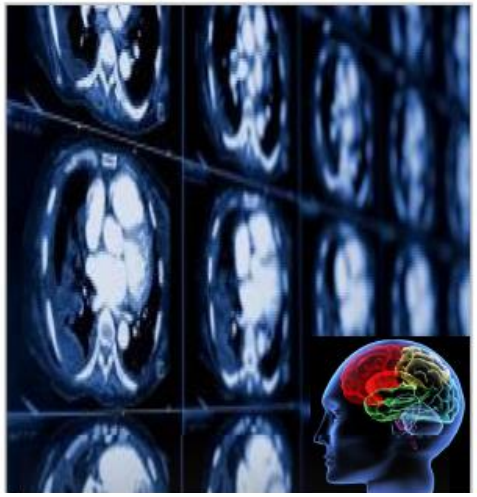
Now it‘s AI doing the driving..

Machine Learning is driving innovation in hardware design

Social media

- Image classification
- Speech recognition
- Language translation
- Language processing
- Sentiment analysis
- Recommendation

Medicine & biology



- Cancer cell detection
- Diabetic grading
- Drug discovery

Media & entertainment

- Video captioning
- Video search
- Real time translation

Security & defense

- Face detection
- Face recognition
- Video surveillance
- Satellite imagery

Autonomous driving

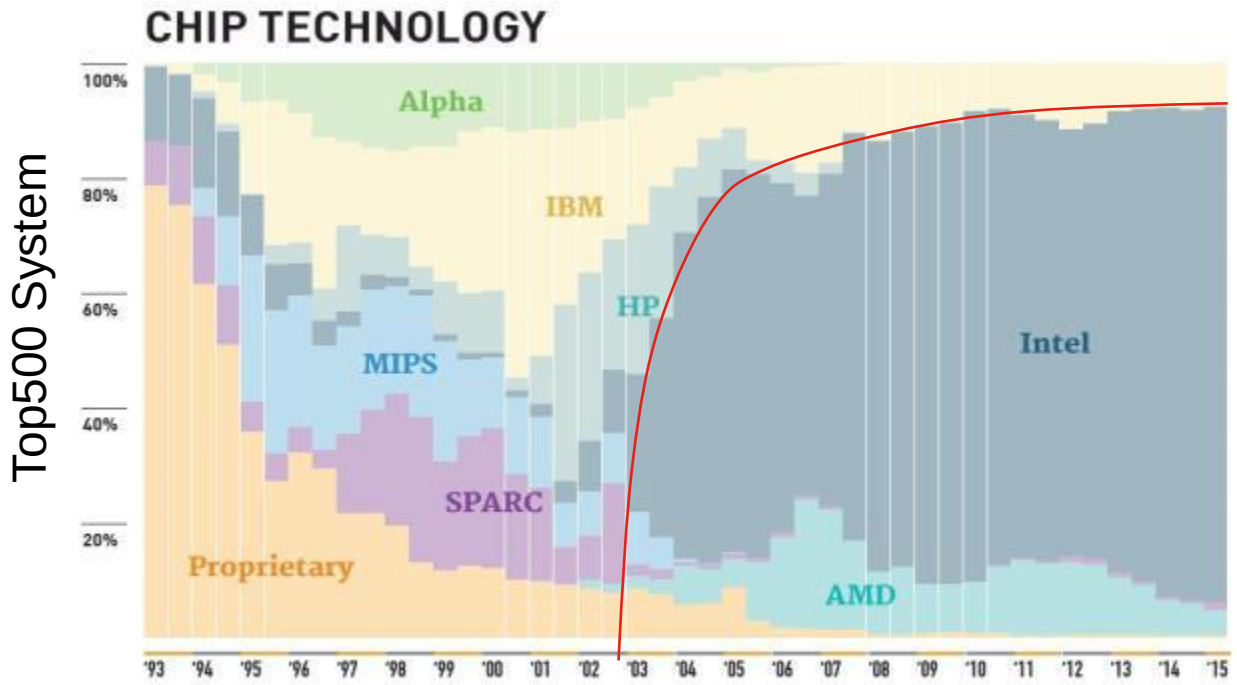
- Pedestrian detection
- Lane tracking
- Traffic sign recognition

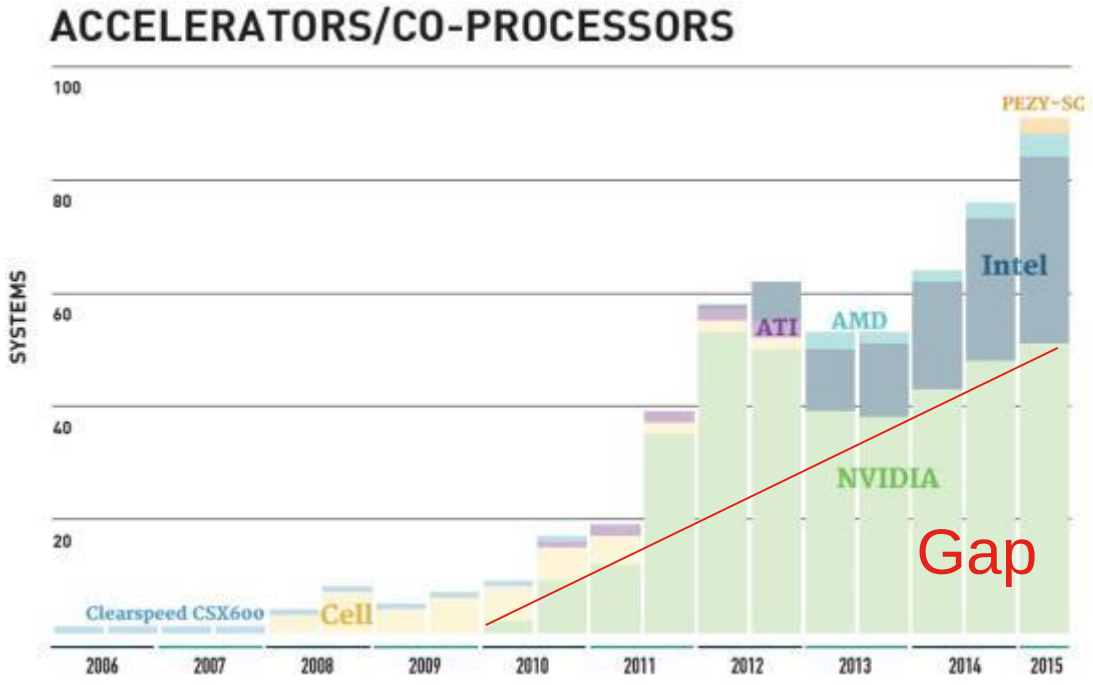
Source: NVIDIA

2016 Lenovo Internal. All rights reserved.

HPC is Evolving – More Specialization. More Choice.

- Intel x86 has dominated the HPC market for most of the 2000's.
- Moore's Law has forced HPC users to embrace hybrid/ heterogeneous technologies
- Non-traditional HPC (i.e. ML, Cloud, Finance, O&G) are leveraging off-loading on FPGAs and GPUs
- Many new processing architectures now viable
- Intel competition desperate to attack dominance = Increases choice and specialization





2016 Lenovo Internal. All rights reserved.

But Which Architecture is Best?

x86	POWER	CUDA	x86 MIC	ARM
				
intel	Power		intel	ARM
AMD		NVIDIA		

2016 Lenovo Internal. All rights reserved.

Clarity in Processing/Technology Choices

	Processing Tech	Approach		Notes
Intel Technologies	**Intel 2S E5**	INVEST & LEAD		Workhorse HPC; Close HPC readiness items
	Intel SMP			Segment unique – e.g. Life Sciences
	Intel 1S	INVESTIGATE		Segment unique – e.g. EDA; need density
	Intel SOC	MONITOR / FAST FOLLOW		Hyperscale, relatively light use in HPC, IoT
	Intel ATOM			IoT
	FPGA			AI/AR/Security – no support in current line
	Xeon Phi (KNX)	INVEST & LEAD		AI/AR + Broad HPC; productize Valiant
Other	**NVIDIA GPU**			AI/AR; add plan for higher density + TTM with Adapters
	NVIDIA NVLink	MONITOR / FF		AI/AR; Potential VLH Co-Sell or Compete with alt tech.
	AMD GPU	ACCOMODATE		Follow NVIDIA plan + Co-Sell Opportunity
	AMD	MONITOR / FAST FOLLOW		Follow portfolio lead on offering
	AMD 1S			Investigate – huge PCIe bandwidth
	ARM 64-bit			Minimize Dev Expense, use STIC/Partners were possible
	Open POWER	COMPETE		

2016 Lenovo Internal. All rights reserved.

HPC Workload Characteristics

Resource contention differs based on workloads

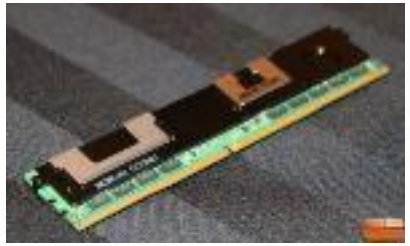
HPC Workload	CPU	Network	Memory	Cache	I/O
Compute-Intensive	✓	✓	✓	✓	✓
Memory-Intensive		✓	✓	✓	
Network-Intensive		✓			✓
Data-Intensive		✓			✓

Question: *Can we design a cost-effective portfolio that can addresses each type of workload without niche products?*

2016 Lenovo Internal. All rights reserved.

Maybe – With New Technology

Storage	Networking	Memory	Processors
• Faster IOPs • Faster concurrent read/writes • High performance • Increased Density • High Capacity • Lower latency	• Lower Latency • Offloading • 10Gb LOMs • RDMA • >=100Gb networks • Core Affinity	• Faster Controller • DDR4 • Flash Storage in memory channel • Higher capacity DIMMs • MCDRAM	• Higher Core counts • SoCs • CPU speeds based on workload • Air or water cooled • High TDP



Question: *Could any one innovation in server technology be a game-changer?*

2016 Lenovo Internal. All rights reserved.

Standards Matter

Lenovo participates in 21 different hardware and software standards organizations.

2016 Lenovo Confidential. All Rights Reserved.

HPC Market Trends and our Strategy

Resurgence of Specialization

Max performance for an expanding set of workloads

Open Everything

Renewed Interest in Open HW and SW Globally

Co-Design is Mandatory

Truly optimized and holistic results based designs

Limited Budgets; Higher Demands

Continued demand for best performance/$ + TCO/ECO/OPEX

2016 Lenovo Internal. All rights reserved.

Modular Design Tenets

SCALE

FLEXIBLE

SIMPLE

Question: Do they hold true for future computing designs?
What's missing?

2016 Lenovo Internal. All rights reserved.

To Tame Complexity - A New Approach is Required

As solutions become more heterogeneous it's critical for us to provide clarity

+

We must get creative on how we bring technology to marketplace

Compete

- Provide alternative
- Partner closely
- Creative approach
- No Touch

OR

Build

- Wide adoption
- High Volume
- Mainstream
- Long Dev Cycle

Buy

- Niche Markets
- Off-Roadmap
- Emerging tech
- Short Dev cycle

Partner

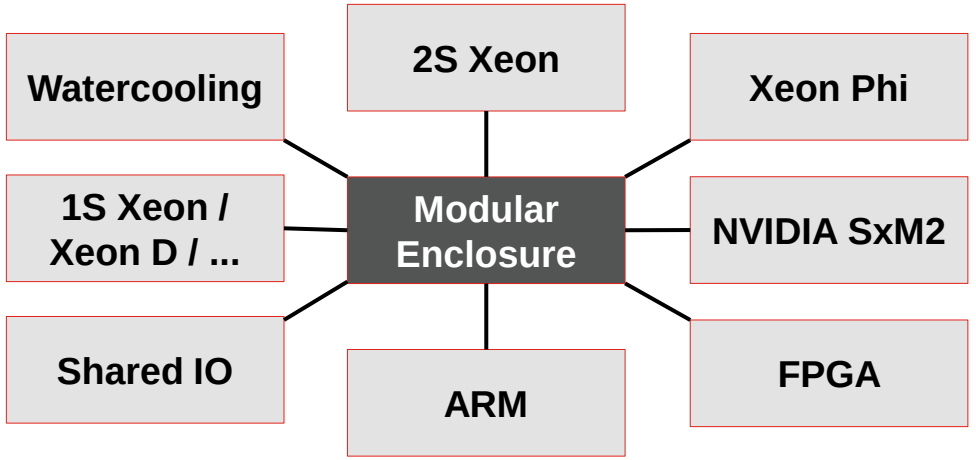
- Co-sell
- Lenovo value add
- Relevance
- Little Development

2016 Lenovo Internal. All rights reserved.

Create a long-term HPC optimal platform

Design Goals

1. **Low Cost** platform for small to large scale
2. **Modular** platform required for broad choice / flexibility
3. **Future proof** platform (KNH, Tinsley, ARM, FPGA, …)
4. **Single** platform for air cooling and water-cooling



Enterprise

Rear IO shuttle for enterprise, converged and general HPC workload

HPC/Specialization

Optimize for front IO using simplified, future proof shuttle for lower cost, increased flexibility

Advantages

- Leverage SoC designs for front IO
- Shuttle contains PSUs, PCMs, Fans, SMM only
- Technology lies in the planar, not in the chassis
- Design once for water and air
- Provides base enclosure for all Dense designs

Further Opportunities

Release Design to OCP
Create and Promote Open Ecosystem
Create a co-design platform with multiple partners

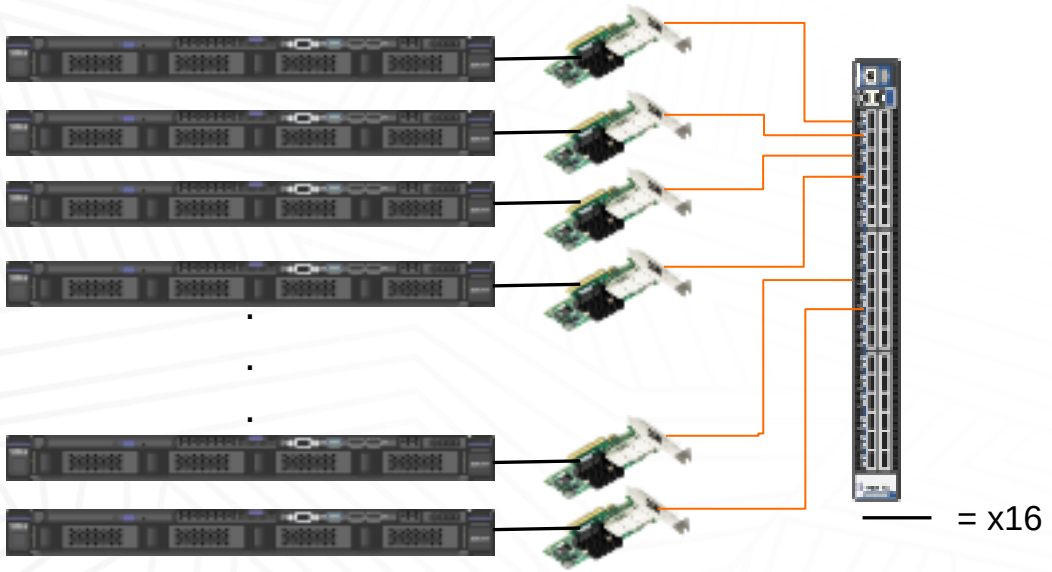
2016 Lenovo Internal. All rights reserved.

I/O Sharing

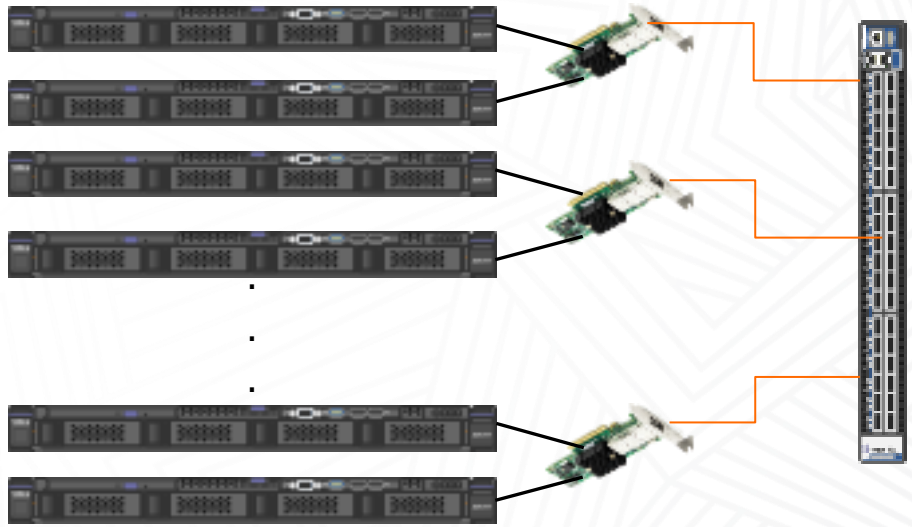
- Networking makes up about 15-20% of cost
- Introduction of 100Gb Ethernet/EDR IB
  - *Very attractive latency*
  - *Pipe is so large many clients will not consume it fully*
  - *Blocking design moves from switch to node*

IT	Qty/Rack	Total Cost	Qty/Rack	Total Cost
PCI Card	72	$36,000	36	$18,000
Cable	72	$7,200	36	$3,600
Ports	72	$14,400	36	$7,200
	TOTAL	$57,600		$28,800

I/O Fixed per Node



I/O Shared between Nodes



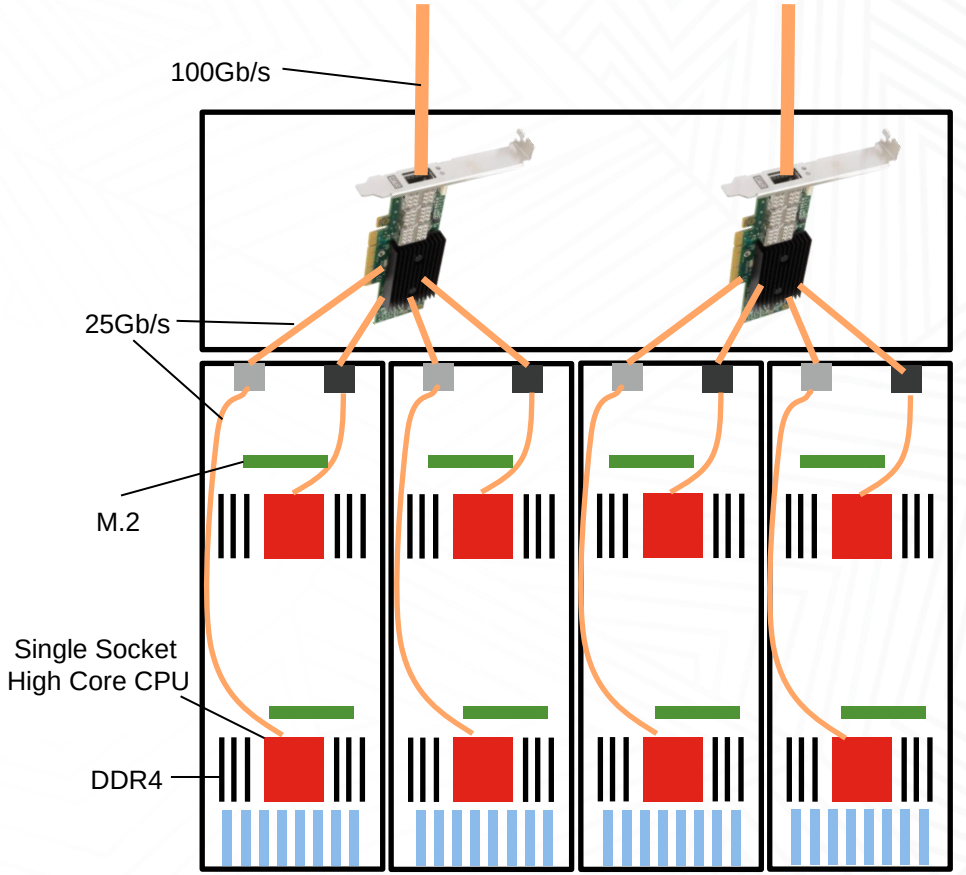
Question: *Is there potential value or other use cases?*

2016 Lenovo Confidential. All Rights Reserved.

Single-Socket I/O Sharing Nodes

Is it time for single sockets?

- ¼ wide trays with 2 x Single Socket servers per tray
- 8 “small drives” per tray in Front, 4 to each single-socket node
- 1 x M.2 per node
- 6 x DDR 4 per node
- 25 Gb/s to node from 1 x 100Gb/s PCIe card via I/O sharing
- 16 sockets in 2U space, 336 sockets in 42U space
- Remove UPI links. Each socket is directly connected to network without dual-rail design.
- Increase density 2x over dual-socket server design



2016 Lenovo Confidential. All Rights Reserved.

Take Ownership of ERE with Lenovo Water Cooling

The Portfolio

Hardware

Optimized server and system design using hot water cooling

Software

Energy-aware management and scheduling E2E

Current Capabilities

<=85% Efficiency at 45°C Inlet temperature
Application profiling and scheduling rules

Future Capabilities

<=95% Efficiency at 50°C Inlet temperature
Energy sentient data center level control

Leading the Industry

How to measure Energy Efficiency

PUE - datacenter efficiency in cooling usage
ITUE - system efficiency in power usage
ERE - datacenter efficiency including reuse heat

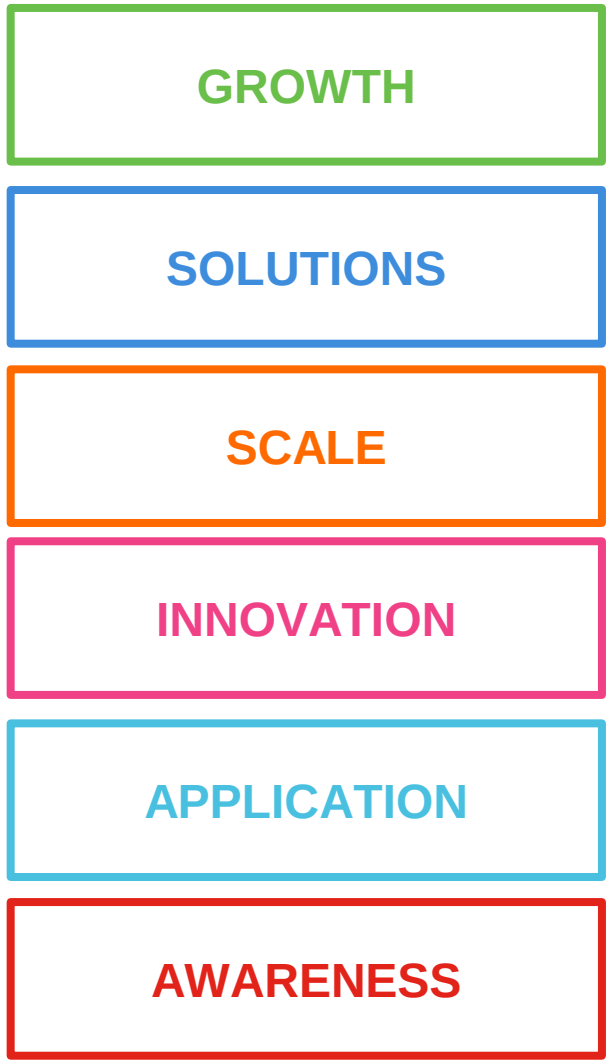
Our Aspiration – Own ERE

1. Maximize heat removal by water
2. Maximize incoming temperature for
   - free cooling all year round
   - efficient use of adsorption chillers
3. Dynamically adjust node power in operation
4. Dynamically control datacenter infrastructure
5. Minimize TCO and lead the industry using

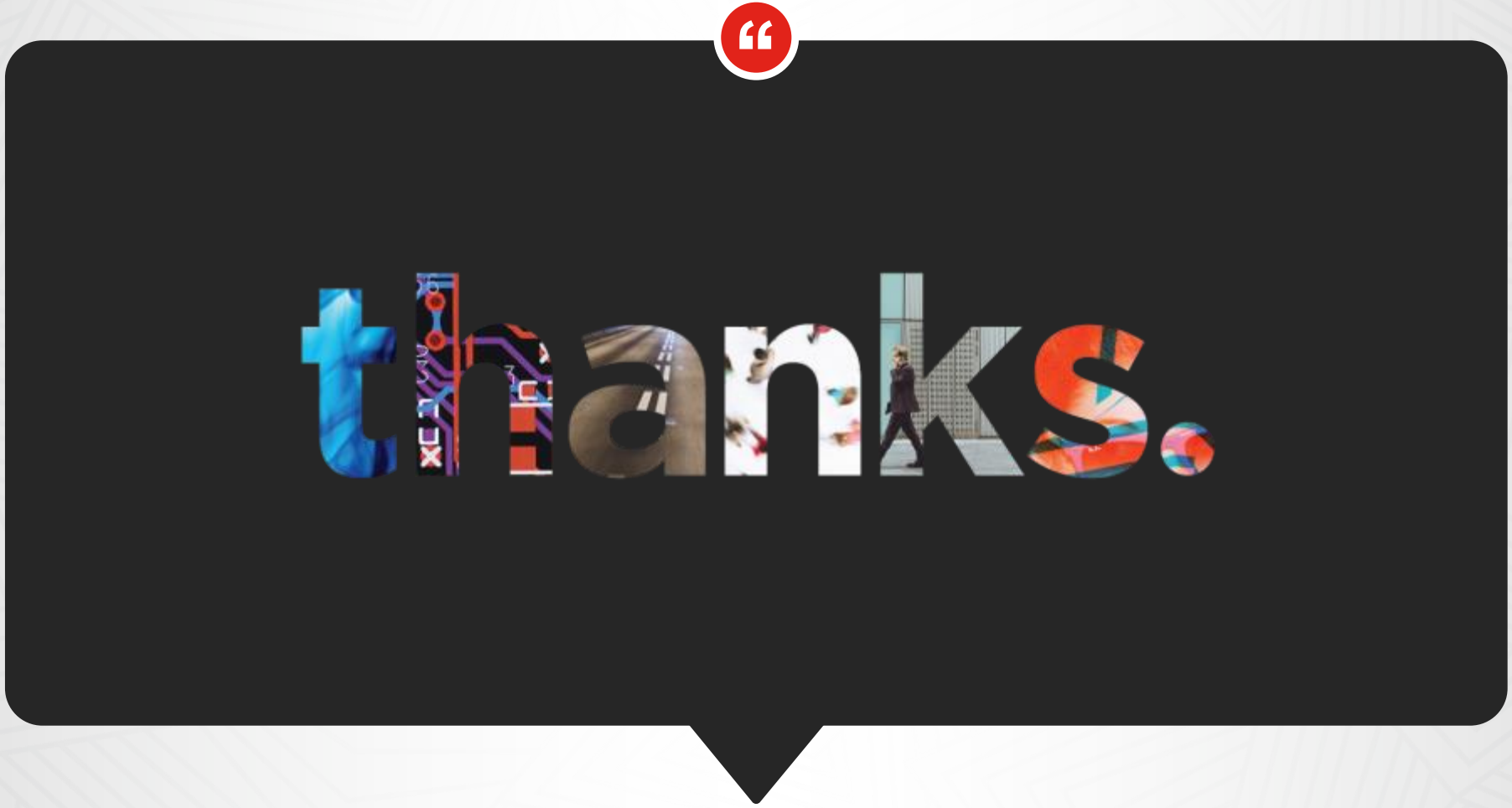
2016 Lenovo Internal. All rights reserved.

Executive Summary

- HPC will continue to be a growth and visibility engine for Lenovo
- Emerging technologies are proving disruptive in the industry with the push towards exascale
- Open Standards will remain and important focus even with the increase in emerging technologies
- Modular designs will become increasingly important
- Cluster designs moving toward more modular fit-for-purpose rather than general purpose

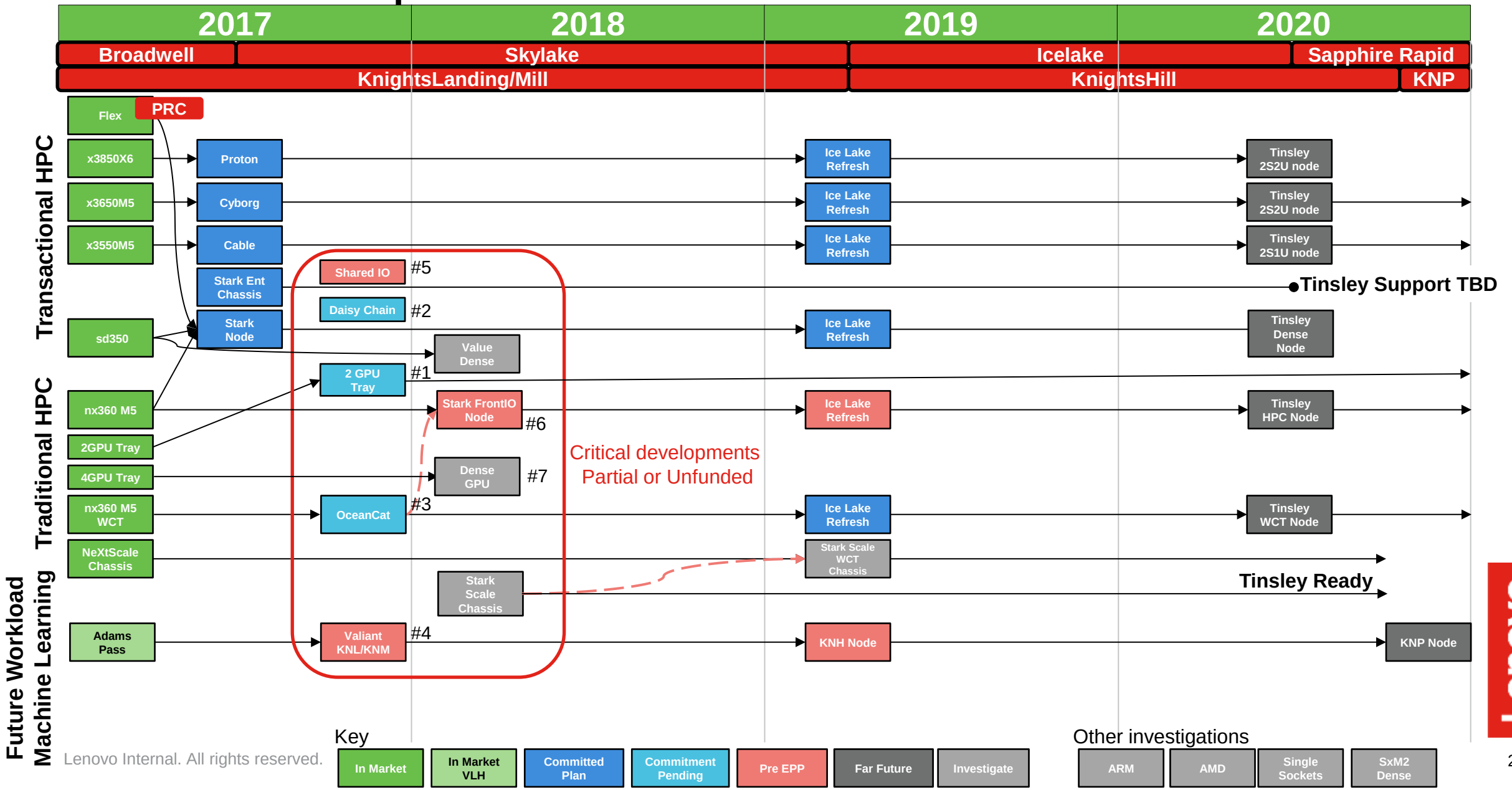


2016 Lenovo Internal. All rights reserved.



Lenovo™

HPC Roadmap Vision



Lenovo Internal. All rights reserved.

System Design Matters for Scale

12 Node

iDataPlex

NeXtScale

Flex

	iDataPlex	NeXtScale	Flex
# of Nodes connected to TOR switch	12	12	0
# of Core SX6036 switches	1	1	0
Total # of SX6036 switches (TOR)	1	1	0
Total # of IB6131 (embedded)	0	0	1
Total # of Racks	1	1	1
Total # of Optical IB Cables	12	0	0
Total # of 10m Optical IB Cables	0	0	0
Total # of IB Cables	12	12	0
Total Number of GbE switches	1	1	1

Key Points:

- Flex solutions are self-contained within a chassis and requires no external switching or cables.
- iDataPlex solution requires use of all Optical cables and rear-facing IB switch
- Flex requires integrated GbE switch. iDPx and NeXtScale can use any TOR.

System Design Matters for Scale

36 Node

iDataPlex

NeXtScale

Flex

	iDataPlex	NeXtScale	Flex
# of connected to TOR switch	36	36	36
# of Core SX6036 switches	1	1	1
Total # of SX6036 switches (TOR)	1	1	1
Total # of IB6131 (embedded)	0	0	1
Total # of Racks	1	1	1
Total # of Optical IB Cables	36	0	0
Total # of 10m Optical IB Cables	0	0	0
Total # of IB Cables	36	36	36
Total Number of GbE switches	1	1	0

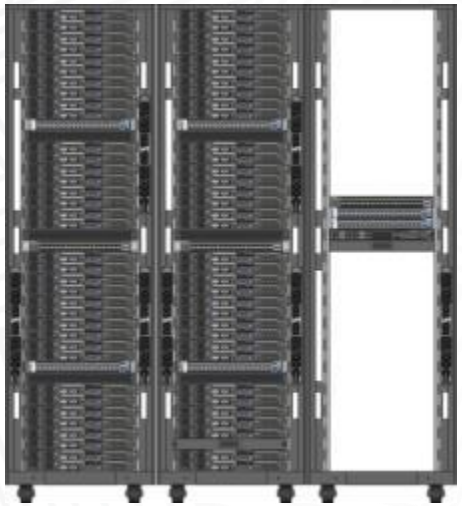
Key Points:

- Flex solution requires 2nd-tier TOR Infiniband switching for non-blocking configurations above 2 chassis.
- iDataPlex fits into a single 42U rack with 36 nodes. Requires all optical cabling and pass-through
- NeXtScale's Block configuration requires specialized brackets to recess switches for copper-only cabling.
- Flex requires C19 Enterprise PDUs

System Design Matters for Scale

72 Node

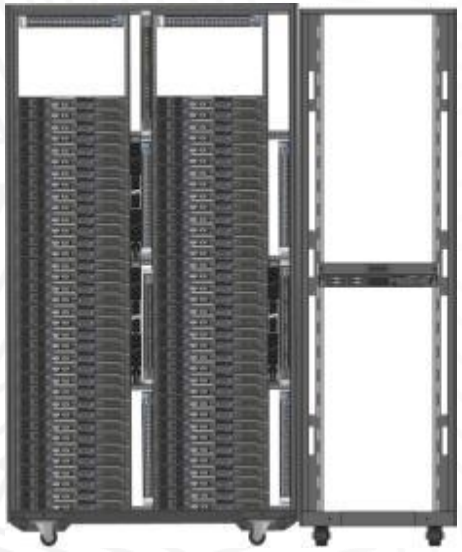
iDataPlex 42U [1]



NeXtScale

Flex

iDataPlex 100U [2]



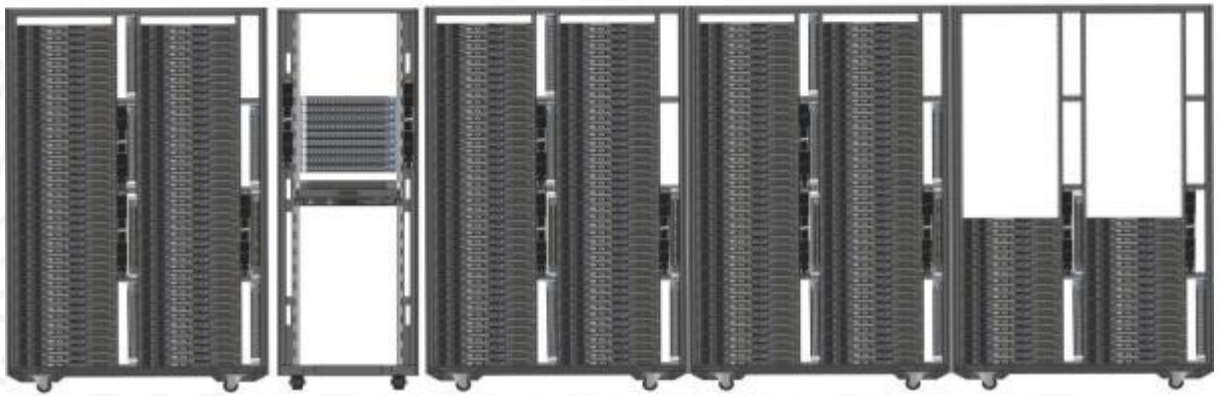
Key Points:

- Flex solution is kept at 12 nodes per chassis to keep chassis' balanced rather than 14 per which leaves last chassis unbalanced in the solution.
- NeXtScale is split between 2 racks to reduce 10m cables by 50%
- NeXtScale requires 10m cables rack to rack as it's hardcoded in x-config.
- iDataPlex 42U rack requires all optical cabling to connect nodes to IB switch

	iDataPlex	NeXtScale	Flex
# of nodes connected to 1st-Tier switch	18	18	0
# of Core SX6036 switches	2	2	2
Total # of FDR switches	4	4	8
Total # of IB6131 (embedded)	0	0	6
Total # of Racks	3[1], 2[2]	2	2
Total # of Optical IB Cables	144[1] , 0[2]	36	36
Total # of 10m Optical IB Cables	72	36	36
Total # of IB Cables	144	144	72
Total Number of GbE switches	3	3	7

System Design Matters for Scale

288 Node



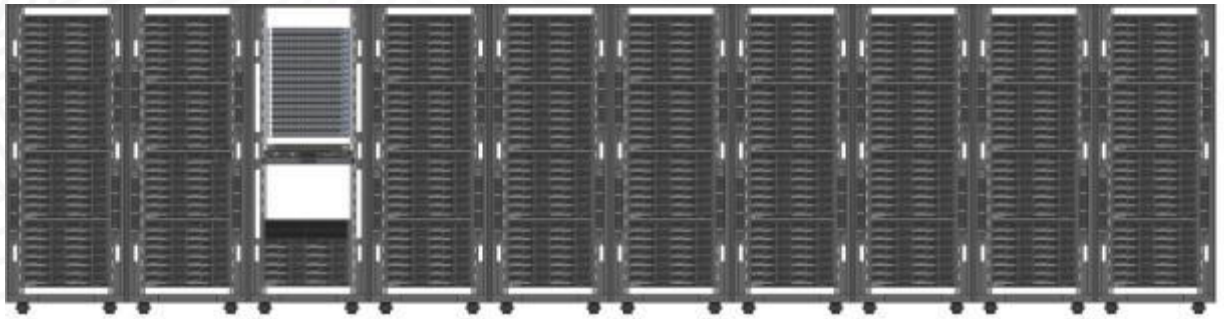
	iDataPlex	NeXtScale	Flex
# of nodes connected to 1st-Tier switch	18	18	0
# of Core SX6036 switches	9	9	12
Total # of FDR switches	25	25	36
Total # of IB6131 (embedded)	0	0	24
Total # of Racks	5	5	7
Total # of Optical IB Cables	300	288	288
Total # of 10m Optical IB Cables	296	270	278
Total # of IB Cables	576	576	288
Total Number of GbE switches	8	9	25

Key Points:

- Flex requires a GbE and IB switch in the chassis driving up switch count.
- For iDPx, x-config was allowed to configure the network resulting in nodes requiring 10m cables to go from rack to 1st-Tier switch.
- All 1st-Tier to core IB cabling is optical
- Flex is 12 nodes per chassis which adds 3 chassis to the configuration but only requires 12 core switches rather than 16 and is balanced.

System Design Matters for Scale

512 Node



	iDataPlex	NeXtScale	Flex
# of nodes connected to 1st-Tier switch	18	18	0
# of Core SX6036 switches	18	18	16
Total # of FDR switches	46	46	53
Total # of IB6131 (embedded)	0	0	37
Total # of Racks	7	8	10
Total # of Optical IB Cables	554	504	576
Total # of 10m Optical IB Cables	514	414	516
Total # of IB Cables	1034	1034	592
Total Number of GbE switches	13	15	38

Key Points:

- Flex's sweet spot config is 504 nodes given the node to switch ratios. Beyond 504, it's impossible to configure it fully non-blocking like NeXt or iDPx
- Remaining nodes are housed in chassis' in the main rack.

HPC Complete Solution Cost Compare – Summary

Extra infrastructure adds to the average cost per node of a system