Hadoop

Hardware sizing – lessons learnt

'The Quest for the ~~Holy Grail~~ Ideal Hadoop Server

Marcel van Drunen
Senior Enterprise Technologist HPC & Cloud
Dell EMEA

3B

People will connect electronically via mobile or Internet technology by 2014

20B

pieces of content shared on Facebook every month

$440B

2011 IT spending in emerging markets - an increase of 10.4% over 2010

2x

IDC estimates that the Digital Universe will double every 18 months

35 ZB

By 2020, the Digital Universe will be 44 times as big as it was in 2009

Outline

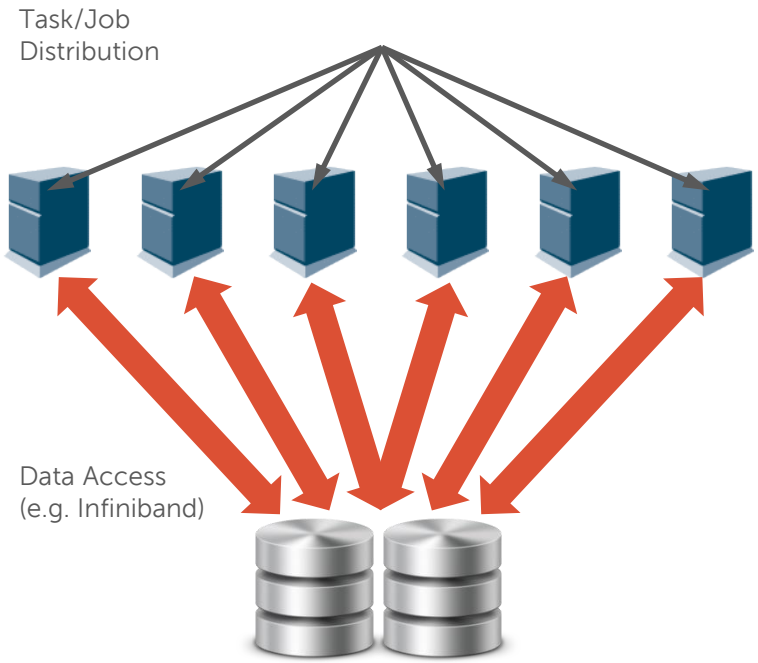
- Intro
- Hardware scaling for Hadoop clusters
- Big Data influences Big Dell

Big Data Infrastructure vs. HPC Clusters

Two Forms of Distributed Computing

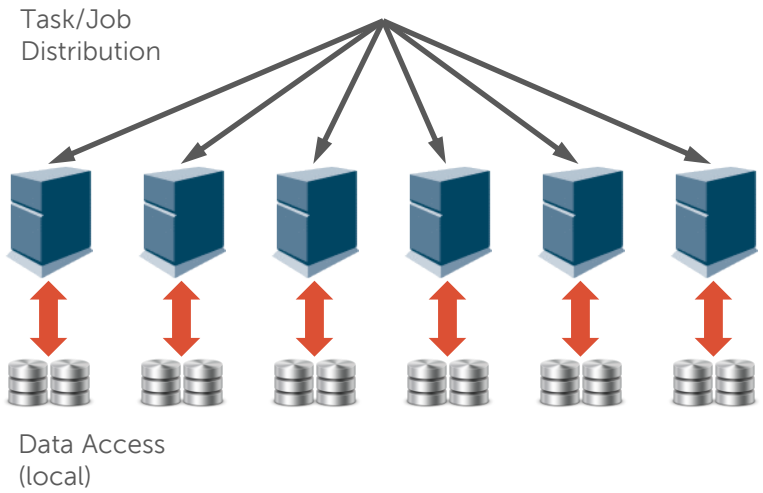
High Performance Compute Clusters

- Parallel File System
  - High throughput
  - All nodes can access all data
  - Compute-centric workloads



Hadoop Cluster Architecture

- Distributed File System
  - Global namespace (ingest!)
  - Nodes just work on local data
  - Data/IO-centric workloads



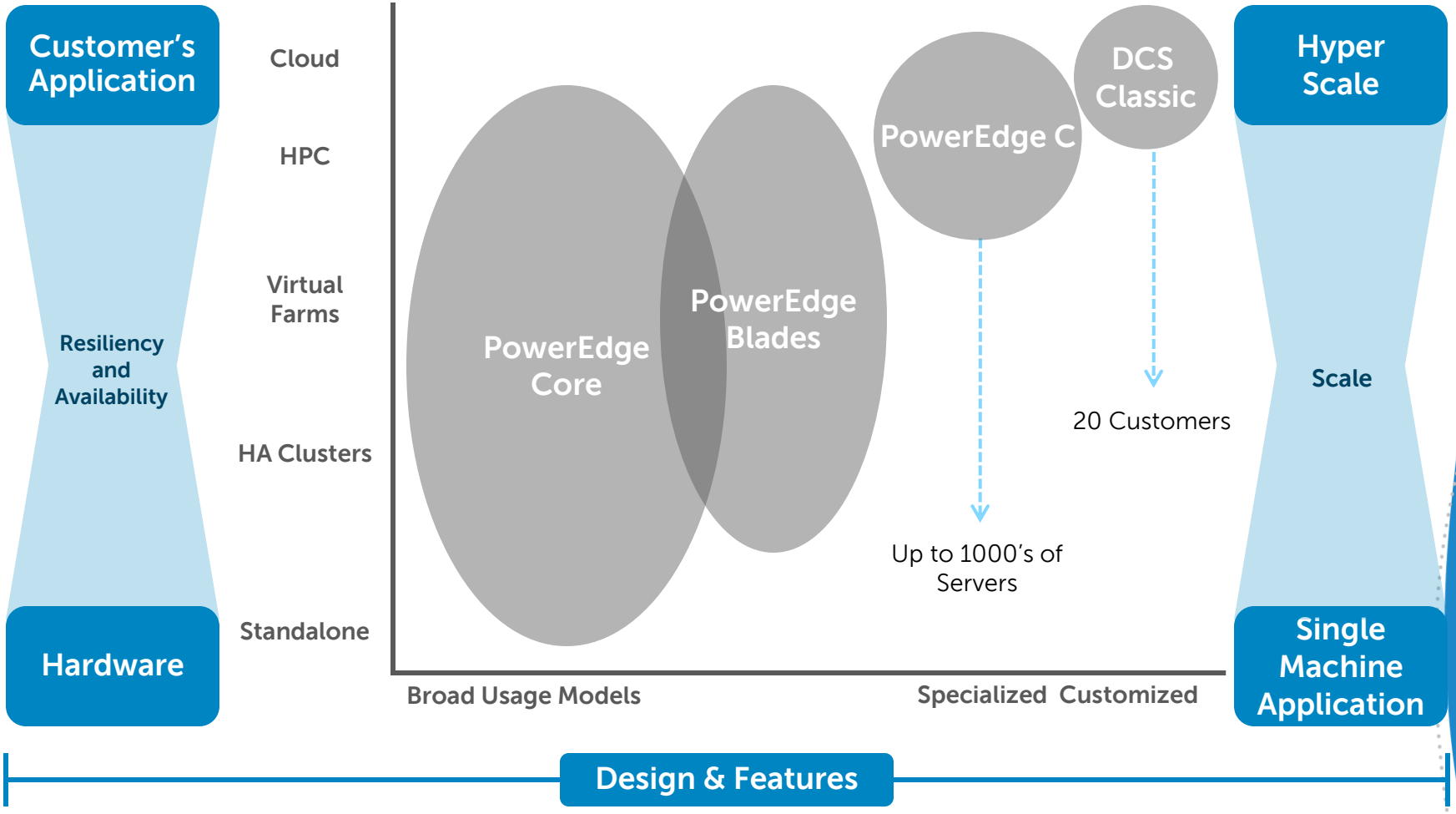
Hadoop & HPC scaling

- Combining Hadoop and regular HPC
  - not a good idea
- Rules of thumb: Data nodes
  - minimum 1 core per disk
  - most workloads HyperTreading counts as second core
  - min. 4 GB per disk, 8 GB is better
- More smaller nodes (12 disks per node maximum)
  - Rebuild times
  - Performance loss during rebuild
  - Network load during rebuild

Hadoop & HPC scaling

- Rules of Thumb: Name Node:
  - 64GB RAM
    - 1GB files → ~13PB of data
    - 128MB files → ~1.3PB of data
    - 1MB files → ~82TB of data
  - Plus OS & JobTracker deamon requirements
  - Plenty of network bandwidth
  - Less nodes, so often overspecced

Big Data influence on Dell portfolio



Please note the diagram is conceptual and not to scale

Dell | Cloudera Solution for Apache Hadoop 2.x

HW + SW + Services		
Hardware	HW Reference Architecture	• PE-R & PE-C Servers* • Storage and compute • PowerConnect switches
	Configuration	• Min of 5 nodes • Deployment guide
Software	Software	• Hadoop Installer (Crowbar) • CDH Hadoop Enterprise • CDH mgmt. applications • Other SW elements installed by Crowbar
	Operating System	• RHEL (if Dell deploys) • Cent OS (if customer deploys)
Services	Deployment	• Onsite HW Install • Onsite SW Install • Whiteboard session & training (via Cloudera)
	Support	• HW: Dell ProSupport • SW: Hadoop support (via Cloudera)

Example: PowerEdge R720xd

- Designed with big data in mind
- Compact 2U form factor
- Capacity, performance, flexibility
- Expansive disk storage

- Quantity:
  - Master nodes: 2 for redundancy
  - Edge node: 1 minimum
  - Data nodes: as many as needed
  - Admin node: 1 minimum

* supported today: C6100, C6105, C2100, C8000, R720, R720XD

Some servers in Dell reference design

- C2100: the server many (DCS) customers ran their first few PB on five years ago

- R720xd: the server they are buying now

Scenario at an unnamed customer

- Wanted single socket nodes
- Four drives per node
- 16 Gb per node

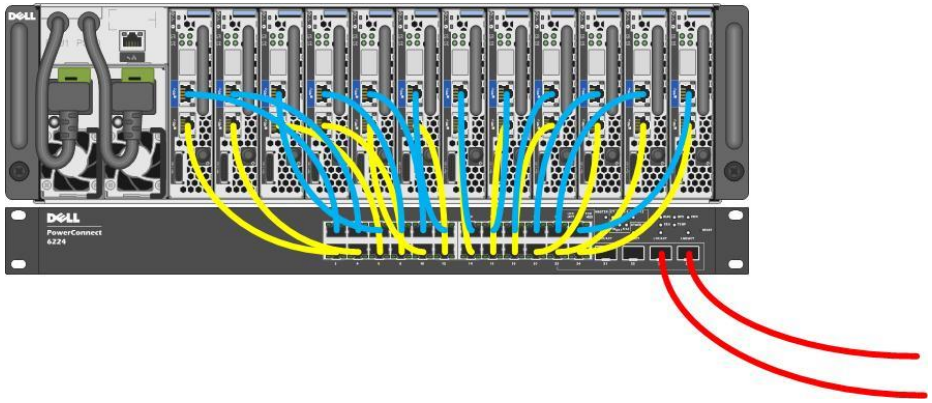
- -> Dell PowerEdge R320
  - › - 39 servers per rack (3 x 13)
  - › - 156 TB per rack raw, 52 TB usable

  - › So what happened when
  - › they wanted a PetaByte?

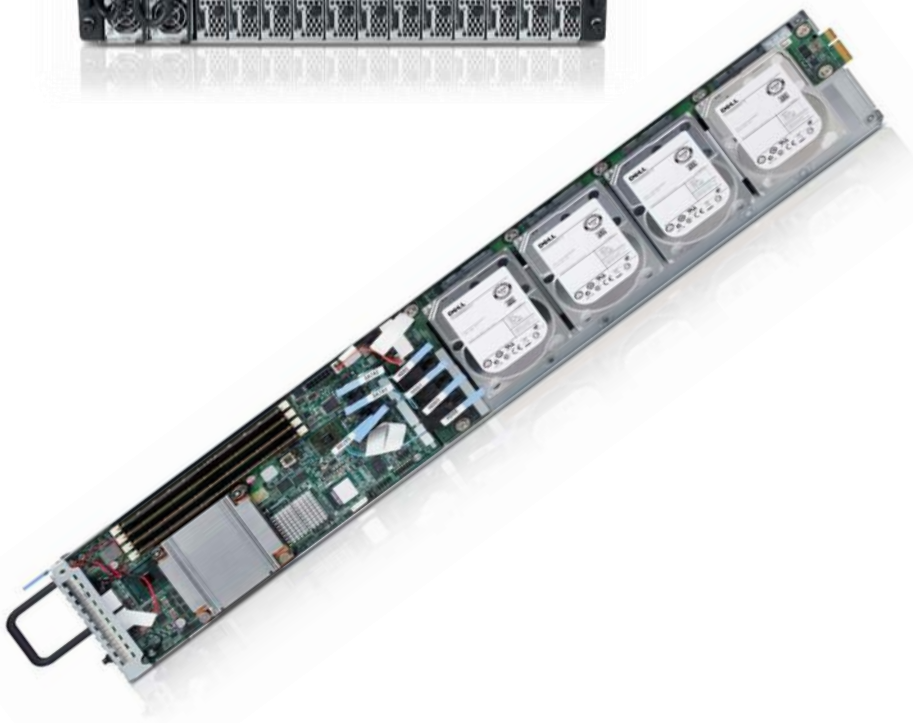
Where density makes sense

- Dell PowerEdge C5220
  - 12 1S servers per 3U
  - Below design with switch per 4U

- 120 nodes per rack, for 480 drives
- That's 3 x 160 TB
- One PB requires 6 racks, plus one for other nodes
- And humongous network bandwidth



I eat my own dog food!

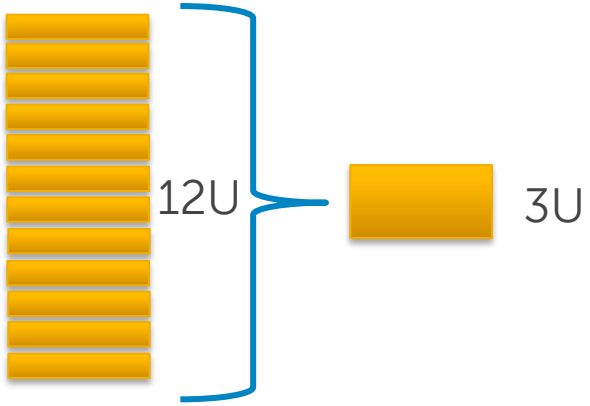


PowerEdge C Microserver Advantage

- **Rare occurrence: a really true marketing slide**

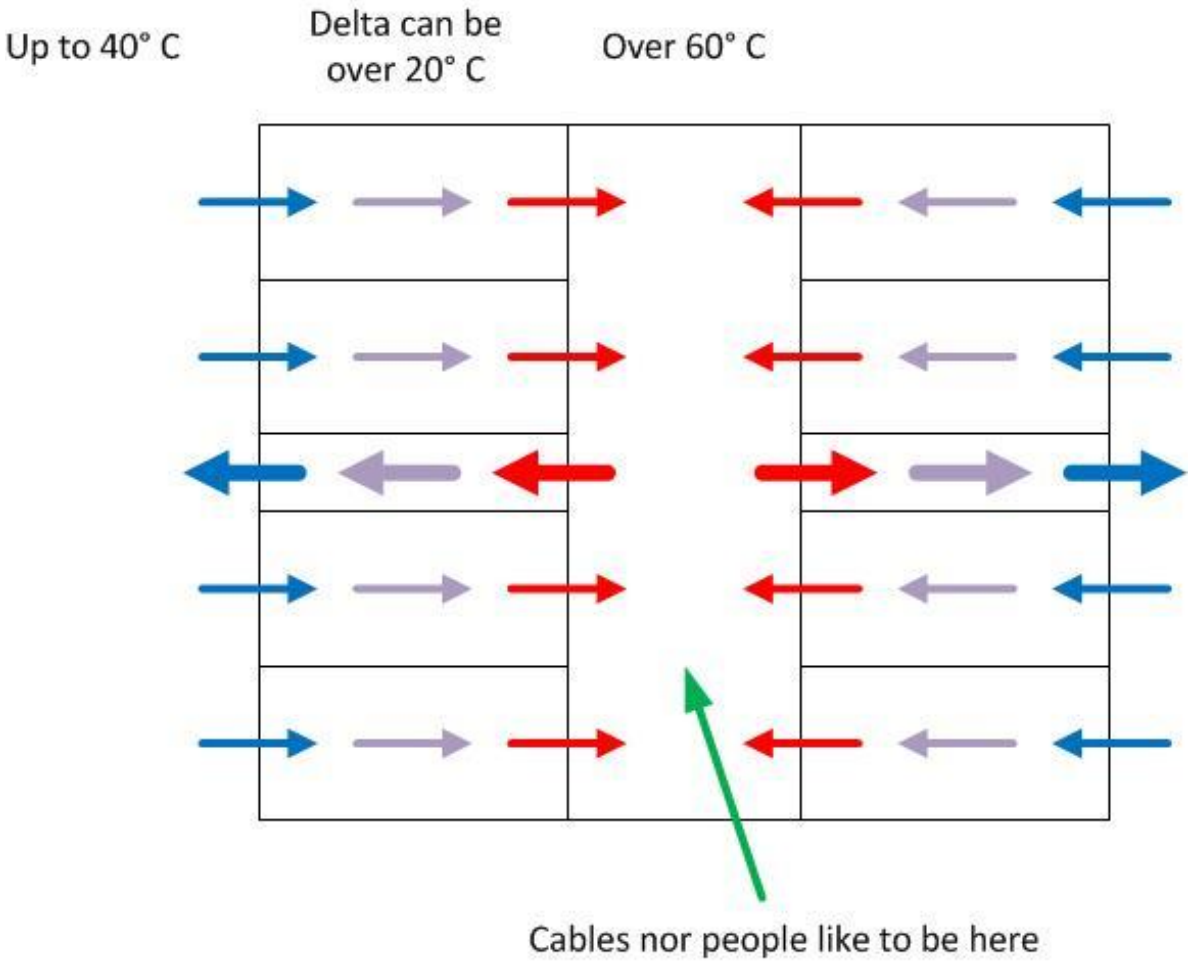
Cooling	Power	Mechanicals

C5220 sled	Comparable 1U
41W	71W

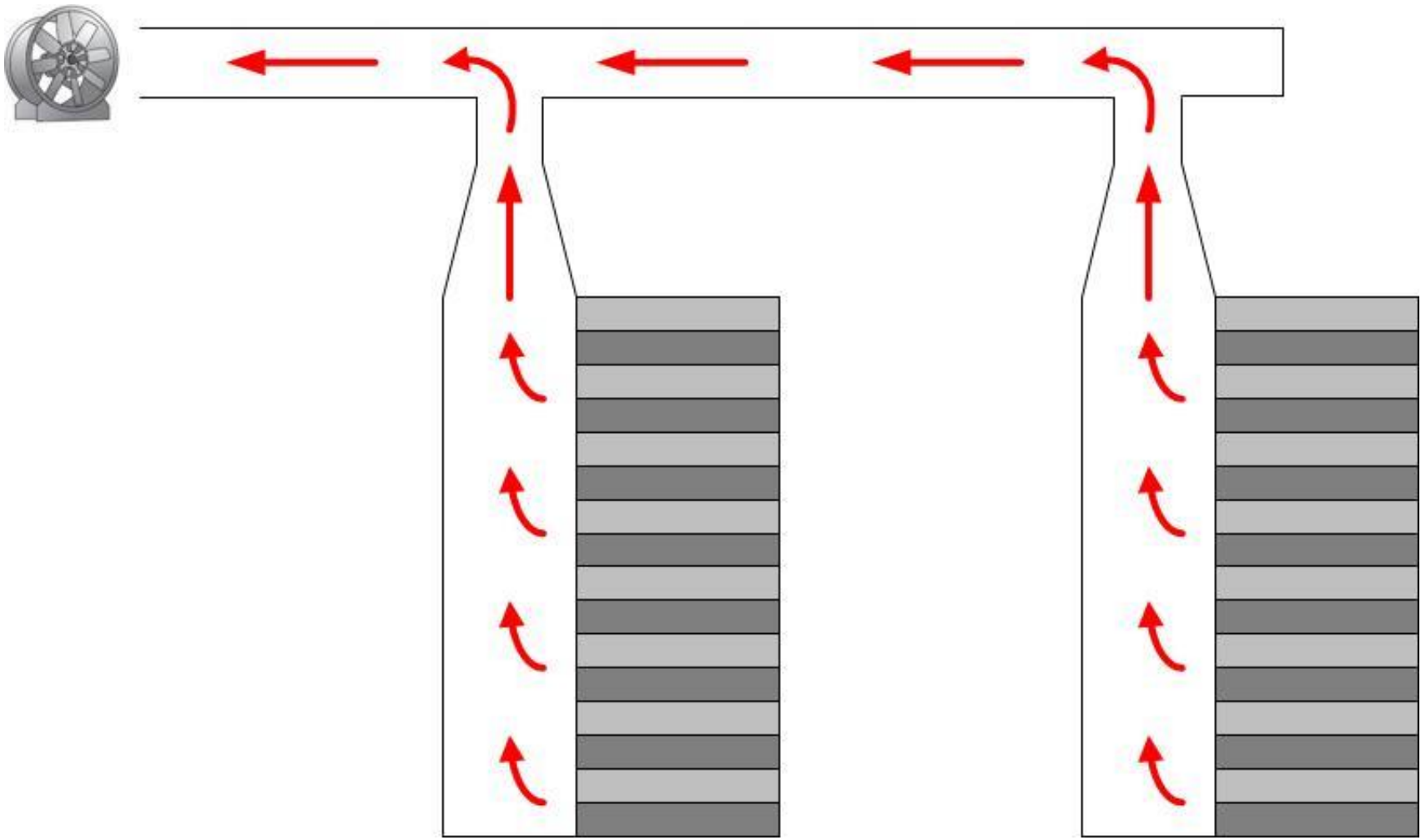


1/6th less amount of fans	40% less power per node	4x the density

The Advantage of Front side serviceability



Cooling with style (at an Italian customer)



PowerEdge C8000 Details

C8000 Chassis
(Internal Power Supplies)

C8000XD Storage Sled

Thank You!

PE-C8220 Single-Width Compute Sled features

Architecture	2S Intel Xeon E5-2600 Series CPUs Intel C600 Chipset
Memory	16 DIMM Slots DDR3 ECC RDIMM Max 256GB 1600MHz
PCI Expansion	1x PCI-E x16 Slot (Single-Width)
Mezzanine Slot	PCI-E x8 Mezz
Drive Controller	Intel C600 (Patsburg)
Drive Bays	2 x 2.5" Internal
HDDs	SAS/SSD/SATA/NLSAS
NIC	Dual Intel 825xx 10 GbE NIC
Management	IPMI 2.0 DCMI 1.0 AST2300 (iKVM) Dedicated Management

PE-C8220X Double-Width Compute Sled features

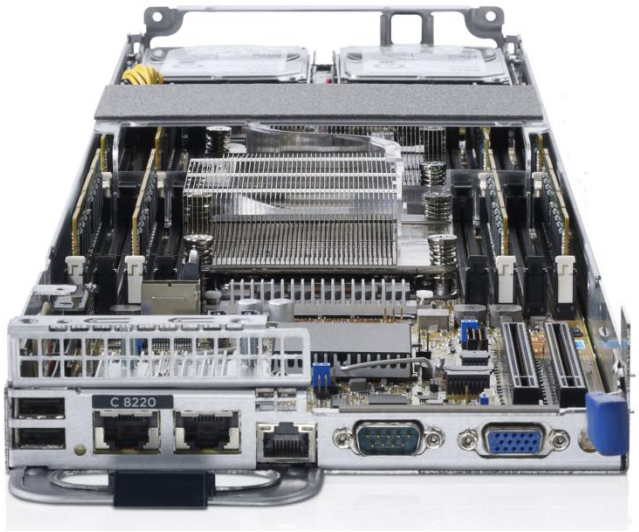
Architecture	2S Intel Xeon E5-2600 Series CPUs Intel C600 Chipset
Memory	16 DIMM Slots DDR3 ECC RDIMM Max 256GB 1600MHz
PCI Expansion	2x PCI-E x16 Slots (Stacked or In-line for GPU Support)
Mezzanine Slot	PCI-E x8 Mezz
Drive Controller	Intel C600 (Patsburg)
Drive Bays	Up to 8 x 2.5" or 4x 3.5" Internal 2x 2.5" External (Full-Width sled only)
HDD's	SAS/SSD/SATA/NLSAS
NIC	Dual Intel 825xx 10 GbE NIC
Management	IPMI 2.0 DCMI 1.0 AST2300 (iKVM) Dedicated Management
GPGPU*	Up to 2 GPGPU controllers

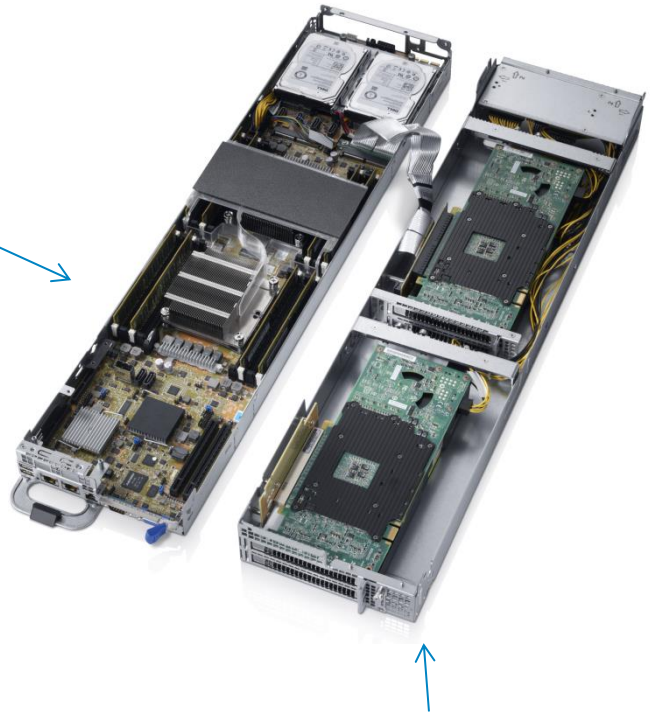
* GPGPU option only available with Double-Width Compute Sled

PE-C8220 and C8220X Comparison

The PE-C8220 and C8220X share the same base chassis. The C8220X expansion holds additional PCI cards, GPU and Hard Drives

C8220 Base





C8220X Expansion sled

PE-C8000 Sled Overview

C8000 Sleds	Specification
SWC Sled C8220	•One Radon MB •Up to 2x 2.5" SATA HDDs •1x PCIE LPX16 add-in controller •1x MEZZ controller
DWC Sled (Non-Hot plug) C8220X	•One Radon MB •Up to 2x2.5" SATA HDDs •Up to 8x2.5" SATA/SAS HDDS(non-HotPlug), or Up to 4x3.5" SATA/SAS HDDs(non-hotplug) •1x LP PCIEX16 + 1xFH PCIEx16 add-in controllers •1x MEZZ controller
DWC Sled (Front Hot plug) C8220X	•One Radon MB •Up to 2x2.5" SAS/SATA Front Hot plug HDDs •Up to 8x2.5" SATA/SAS HDDS(non-HotPlug), or Up to 4x3.5" SATA/SAS HDDs(non-hotplug) •2x LP PCIEX8 add-in controllers •1x MEZZ controller
DWC Sled (GP GPU) C8220x + GPU	•One Radon MB •Up to 2x GPGPU controllers •1x MEZZ controller
DWS Sled C8000XD	•LSI SAS Expander with 2 x LSISAS2x28 chips •Up to 12 x 3.5" SATA/SAS HDDs hotplug •Up to 12 x 2.5" 15mm SATA/SAS HDDs hotplug •Up to 24 x 2.5" 9.5mm SATA/SAS HDDs non-hotplug
Power Sled	•2 x 1400w Delta Platinum PSU •Up to 2 PSU sleds in system

PE-C8220

PE-C8220X

PE-C8000XD