Server technologies in the light of accelerated computing

Dr. Roland Kunz

7.11.2023

Top strategic technology themes through 2025

SYSTEM OPTIMIZING TECHNOLOGIES

New processors, persistent media, fabrics and data center architectures

CLOUD STACK EVOLUTION

Technologies that enable intelligent workload placement across different clouds, the development of cloud native apps and disparate clouds to work as an aggregated system

EDGE & DECENTRALIZED IT

Technologies, system designs, frameworks and security and management tools that drive the creation of edge-centric architectures and software

DATA MANAGEMENT

Tools and technologies that enable a holistic approach to the data lifecycle; e.g. metadata lifecycle management, automated data catalog and data-as-a-service

DATA SCIENCE (AI & ANALYTICS)

Analytics and AI/ML technologies that address the growing needs of data scientists and the ecosystems they leverage

INTRINSIC TRUST & SECURITY

Technologies and use cases enabling security to be built into all the components and layers of a solution in an increasingly automated way for foolproof, scalable and end-to-end protection of modern, distributed architectures

NEXT GENERATION COMMUNICATIONS

New high-performance wireless, wired and virtualized technologies to connect Things at the Edge and Apps across the multi-cloud

INTELLIGENT AUTOMATION & ORCHESTRATION

Machine learning and analytics embedded into systems combined with Automation/Orchestration systems to enable self-driving, self-optimizing and auto configuring infrastructures and systems

CITIZEN DEVELOPERS & DEVOPS

Technologies, frameworks and toolchains that democratize and automate application development and drive innovation from across an enterprise

AUGMENTATION

Comprises Augmented Perceptions, Interactions and Cognition and the underlying systems that enable them

SUSTAINABILITY

Emerging technologies and strategies that embrace and enable sustainable products, circular economy, energy efficiency and waste reduction

DELL Technologies

Partnering on the path to a **green data center**

ENERGY EFFICIENT HARDWARE
Dell's data center solutions are designed to deliver high performance per watt

PLATFORM POWER MANAGEMENT
Dell servers have built In BIOS And iDRAC settings to help reduce energy waste

WORKLOAD MIGRATION
Dell solutions can help customers manage workloads on premise and in the cloud

RESPONSIBLE RETIREMENT
With Dell recovery and recycling services customers can retire equipment responsibly

DC POWER MANAGEMENT
OME power manager delivers telemetry to help lower customers carbon footprint

OPTIMIZED THERMALS
Dell designs hardware with optimized cooling and power capabilities

DELL Technologies

Reducing the carbon footprint of your IT hardware

Maximizing energy efficiency is critical to lowering Product Carbon Footprint in the data center

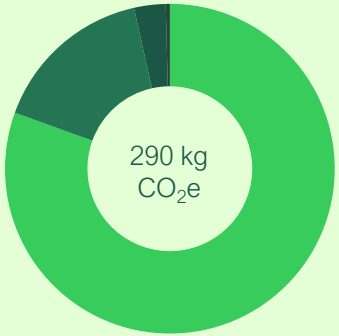
Key focus areas for reduction are:

1. **Energy** (Use)
2. **Materials** (Manufacturing)
3. **Packaging** (Transportation)
4. **Repairability & Upgradability** (Manufacturing & End-of-Life)
5. **Reuse & Recycling** (Manufacturing & End-of-Life)

Latitude 5430[1]

Product Carbon Footprint (PCF)

- Manufacturing 80.6%
- Use 16.0%
- Transport 3.1%
- End-of-Life 0.4%

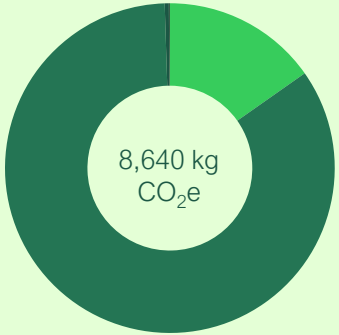


*Based on most commonly sold configuration vs. lowest configuration**

PowerEdge R740[1]

Product Carbon Footprint (PCF)

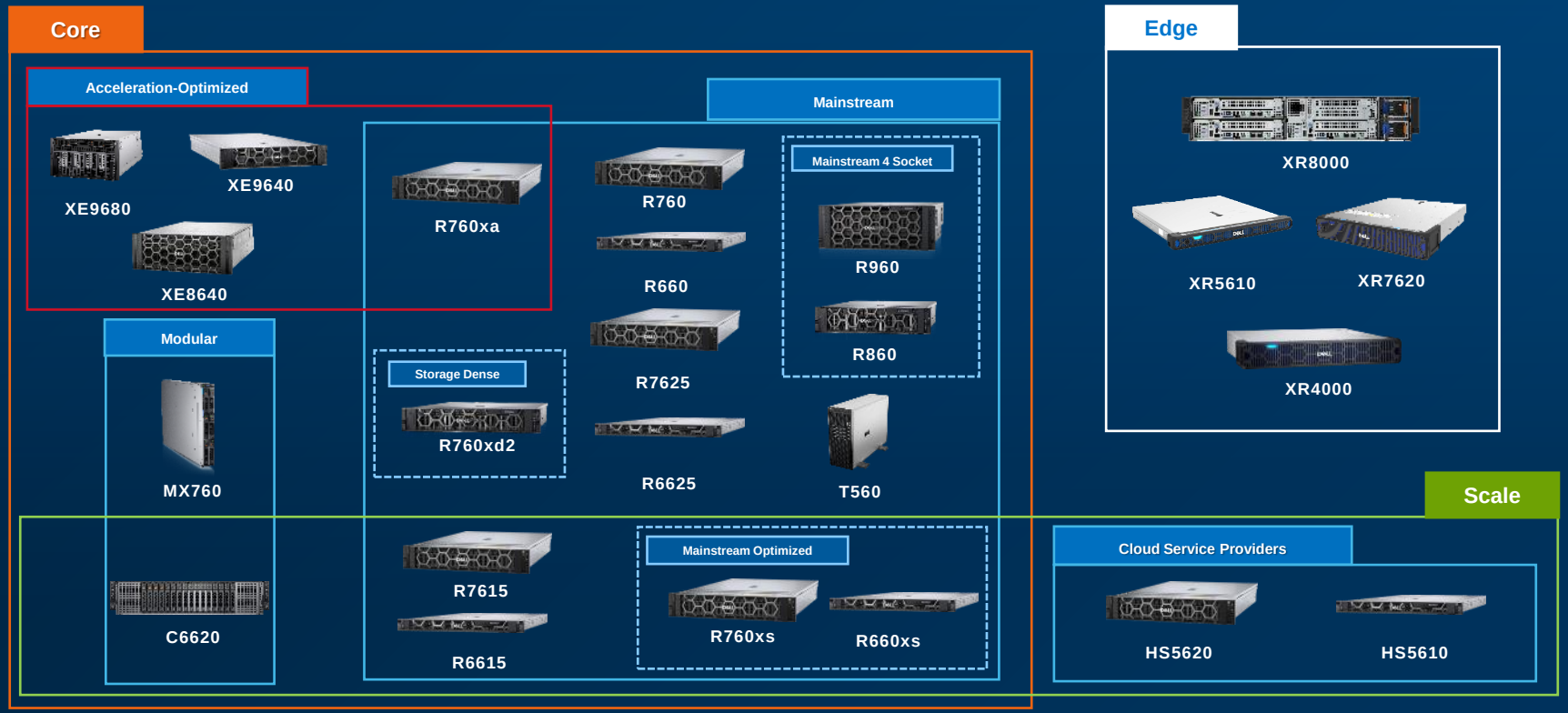
- Manufacturing 15.2%
- Use 84.0%
- Transport 0.4%
- End-of-Life 0.1%



*Based on most commonly sold configuration vs. lowest configuration**

*Results may vary based on region and energy mix

PowerEdge portfolio 2023



Industry Enabled Technologies Overview

Next Generation Intel & AMD Processors

- Intel 4th Gen Xeon (Sapphire Rapids)
  - ✓ Up to 60 cores/CPU*
  - ✓ 50% performance increase over Ice Lake
- AMD 4th Gen EPYC (Genoa)
  - ✓ Latest 5nm technology with up to 96 high-performance “Zen 4” cores
  - ✓ 1.5X & 1.25X the density and power over Milan

Memory: DDR5

- DDR5 (4800MT/s)
  - ✓ Latest DRAM technology with higher speed & bandwidth
  - ✓ Greater efficiency with 2 channels per DIMM
  - ✓ Improved RAS features with on-die ECC
  - ✓ Lower power
  - ✓ Enhanced telemetry for temperature reporting and systems management

PCIe Gen5 Capability

- Doubles throughput compared to PCIe Gen4
  - ✓ Benefits NVMe drives, GPUs, and some networking cards

EDSFF E3.S NVMe Gen5

- E3.S form factor will be introduced with PCIe Gen5 NVMe drives
  - ✓ Benefits density, thermals, and improved packaging in space constrained servers
- Double the performance over NVMe Gen4

*Max 60 cores for 4S CPUs, max 56 cores for 2S CPUs

DELL Technologies

Dell enabled Technologies Overview

Next Gen HWRAID (PERC12)

- New gen controller with 2X better performance over PERC11 and 4X better than PERC10
  - ✓ Supports all drive interfaces: SAS4, SATA & NVME
  - ✓ x16 connectivity to devices to take full advantage of PCIe Gen5 throughput

System Cooling & Efficiency

- Power Manager & Smart Cooling
- High Power Optimized Airflow chassis design to maximize air cooling capabilities
  - ✓ Support for XCC/HBM in air-cooled chassis
- Optional CPU direct liquid cooling (DLC) solutions

BOSS-N1

- Segregated RAID controller for OS with secure UEFI boot that is rear facing and hot-pluggable
  - ✓ Enterprise-class 2 x M.2 NVMe devices with strong endurance and high quality that provide increased performance over BOSS-S1 with SATA drives

Data Processing Unit (DPU)

- SmartNIC with hardware accelerated networking and storage that enables customers to save CPU cycles
  - ✓ Improved security, running workloads and security software on different CPUs (“air gap”)
  - ✓ Offload hypervisor, networking stack, and storage stack to the DPU making them OS independent

System Management

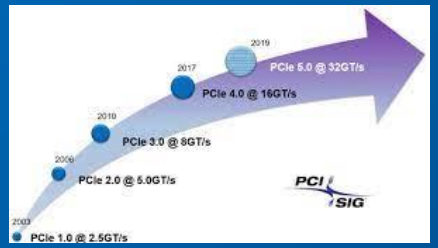
- Seamless integration of new 16G servers into your existing processes and tool set
- Complete iDRAC9 support for all components
  - ✓ PERC12, BOSS N-1, PCIe Gen5 devices, UEFI Secure Boot, Smart Cooling, DPU’s, and more

Security

- TLS 1.3 with FIPS certification, SEKM 2.0 with support for NVMe drives and VxRail
- End-to-end threat management with Zero Trust approach
  - ✓ Silicon-based platform root of trust, multi-factor authentication (MFA), inventory and platform component tracking during delivery, tamper protection during shipping

Future technology disjunction

Power consumption of current CPU and GPU gen increases massively (>350 / 500 Watt TDP)
Cooling will increase and need new technology (non-air cooled)
New AI and ML applications will eat up those resources



Power and Cooling of existing Datacenters almost stays the same and is often limited to ~10-15 kw/rack*
Density is not longer possible with legacy environments

Compute power required is not increasing massively for existing workloads
New methods of power management might mitigate some of the requirements
DPUs can perform some tasks at lower power consumption

Cooling

Our world class engineers designed PowerEdge servers for ultimate thermal performance.

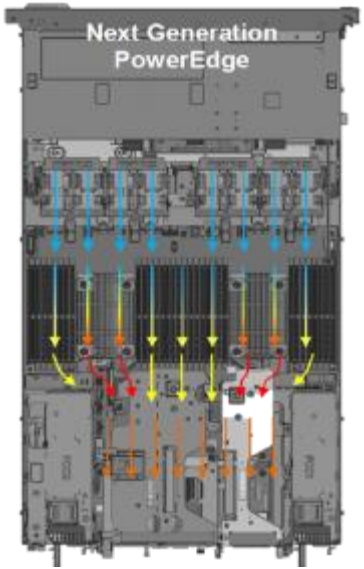
With a new layout and high-performance fans, hot air exits the system quickly and efficiently.

- Latest Intelligent thermal algorithms minimize fan and system power consumption while maintaining component reliability
- Enables custom cooling options that can be managed via iDRAC GUI

3rd generation DLC solutions enable dense configs with high TDP CPUs

- expanding to cover more platforms, with solutions backed by Dell Services
- New 2U 4-way DLC-cooled GPU system in CY23

PowerEdge Smart Cooling Solutions



Overview

- Next generation technologies are driving power and heat higher and higher
- PowerEdge ensures no-compromise system performance through innovative cooling solutions, while also offering customers options that fit their facility or usage model needs (one size does not fit all!)

Air Cooling

- 16G delivers innovations that extend the range of air-cooled configurations
- **Advanced designs** - airflow pathways are streamlined within the server, directing the right amount of air to where it's needed
- **Latest generation fan and heat sinks** – to manage the latest high-TDP CPUs and other key components
- **Intelligent thermal controls** – automatically adjusts airflow during workload or environmental changes, seamless support for channel add-in cards, plus enhanced customer control options for temp/power/acoustics

Direct Liquid Cooling (DLC)

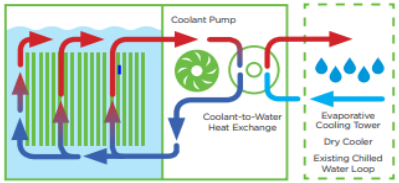
- For high performance CPU & GPU options in dense configurations, Dell DLC effectively manages heat while improving overall system efficiency
- DLC options available for C-series, select R-series, 4S and MX platforms
- New: purpose-built liquid-cooled 2U 4-way GPU accelerator system

Edge Cooling

- New XR edge platforms deliver performance with extended temperature range support from -5ºC to 55ºC

DELL Technologies

Cooling Technology Comparisons

	Air cooling	Air + Supplemental	Direct Liquid Cooling (DLC)	Immersion
Cooling Solution Options		 + 		
Products	• Traditional air-cooling & air-handling equipment • Containment	• In-row coolers • Rear Door Heat Exchangers (RDHx) • Containment (hot & cold aisle)	• CPU/GPU Cold-plate loops • Rack/facility level DLC products required	Single-phase (1P) and Two-phase (2P) Immersion tank solutions
Environments	Traditional data centers	Traditional data centers, with facility water	Traditional data centers, with facility water	• Non-traditional spaces, no conditioned air required (ex. - warehouse) • Note: facility water required
Main usage model	• Low to Mid-density racks • Up to ~ 15kW/rack	• **Mid to High-density racks** • Up to ~30kW/rack	• **Systems with high TDP parts** • High-density racks, up to ~80kW/rack	• **Limited/no air cooling available** • High-density racks, or high TDP parts
Typical Cost Adder	*NA*	+	+ +	Single phase (1P): + + Two-phase (2P): + + +
Availability	Standard cooling	Standard server cooling + 3rd party supplemental cooling solutions	Dell factory supported configurations	Dell OEM project engagement

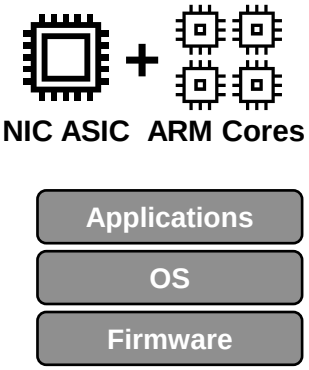
DELL Technologies

Data Processing Unit

(DPU aka SmartNIC)

- **Save CPU cycles with hardware accelerated networking and storage**
- **Improve security by running workloads and security software on different CPUs ("air gap")**
- **Offload hypervisor, networking stack, and storage stack to the DPU making them OS independent**
- **Enable landlord/tenant models by isolating tenants not just with software, but also through hardware**

DPU Definition



- DPU is a combination of ARM Cores and a NIC ASIC
  - ARM cores run an OS and applications
  - NIC ASIC has hardware accelerate networking and storage
- PCIe form factor only

VMware ESXi 8.0 Distributed Services Engine on DPUs

(formerly VMware's Project Monterrey)

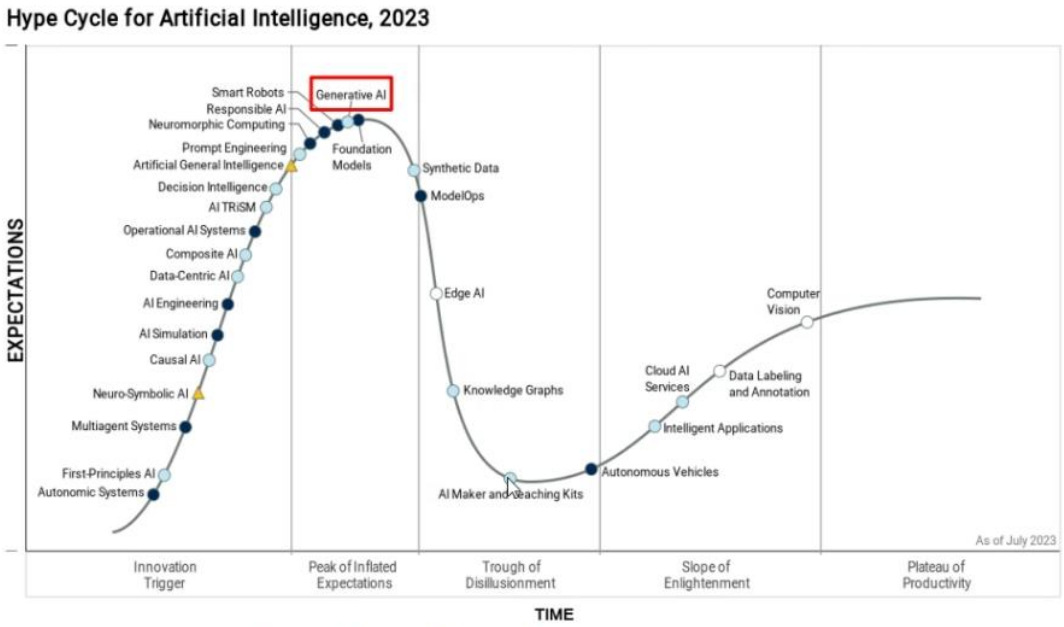
- PowerEdge servers will support VMware ESXi running on a DPU
- These DPUs will be fully integrated into PowerEdge systems management - DRAC, OMIVV, and OME
- This solution will be supported with VxRail
- This solution has special hardware integrations
  - A cable that provides a serial connection as well as a high-speed connection to the iDRAC (same type of connection that a LOM has)
  - In 16G support for "Always On" where the DPU can be powered on and off independently from the server. This is necessary for the DPU "landlord-tenant" model

NVIDIA Channel DPUs

- PowerEdge supports NVIDIA channel DPUs that will run Linux
- Channel DPUs will have limited systems management integration (i.e., the server cools the DPU)
- Channel DPUs will not support VMware ESXi

Accelerated portfolio

Gartner Hype Cycle



Gartner

PowerEdge.Next GPU Acceleration Server Portfolio

PCIe Optimized	4-way SXM	4-way Dense	8-way SXM
			
R760XA	**XE8640**	**XE9640**	**XE9680**
• **2U** monolithic • **2-socket** Sapphire Rapids CPU • **Up to 4 x** double-wide GPUs • **Up to 12 x** single-wide GPUs • **Full PCIe GPU portfolio** supported • **Air cooled** with optional liquid cooling for CPU	• **4U** monolithic • **2-socket** Sapphire Rapids CPU • **4 x Nvidia H100 SXM** NVLink GPUs; • **Air cooled**	• **2U** monolithic • **2-socket** Sapphire Rapids CPU • **4 x Nvidia H100 SXM** NVLink GPUs (Q3 availability); -or- • **4 x Intel Data Center Max 1550 OAM** XeLink GPUs (Q2 availability) • **Direct liquid cooled** CPUs and GPUs	• **6U** monolithic • **2-socket** Sapphire Rapids CPU • **8 x Nvidia H100 SXM** NVLink GPUs -or- • **8 x Nvidia A100 SXM** NVLink GPUs • **Air cooled**
High performance 2U server purpose built for dense PCIe GPU acceleration. Maximize AI, HPC, VDI and performance graphics supporting multiple GPU choices. Use cases: • AI/ML Inferencing • AI/ML Training • Rendering/Perf. Gfx • VDI	Accelerate and automate analysis into insights. Maximize AI initiatives performance in a 4-way GPU, 4U server. Use cases: • AI/ML Training • HPC Modeling & Simulation	Push performance boundaries with a dense form-factor, liquid cooled approach to AI initiatives. Smallest form factor 4-way GPU, dense 2U AI/ML/DL & HPC server. Use cases: • AI/ML Training • HPC Modeling & Simulation	Modernize operations and infrastructure to drive new AI initiatives. Optimized for demanding AI/Machine Learning & Deep Learning applications Use cases: • Large AI/ML/DL Training

DELL Technologies

NVIDIA GPU portfolio

	H100			A100		A30	L4	A2	L40S	L40	A40	A10	A16
Design	Highest Perf AI, LLM, HPC, DA			High Perf Compute		Mainstream Compute	Universal AI, Video, and Graphics	Entry-Level Small Footprint	Gen AI	Powerful Graphics + AI	High Perf Graphics	Mainstream Graphics & Video with AI	High Density Virtual Desktop
Form Factor	SXM5	x16 PCIe Gen5 2 Slot FHFL 3 NVlink Bridge	X16 PCIe Gen5 Dual 2 Slot FHFL using 3 NVLink Bridges	SXM4	x16 PCIe Gen4 2 Slot FHFL 3 NVLink Bridge	x16 PCIe Gen4 2 Slot FHFL 1 NVLink Bridge	X16 PCIe Gen4 1 slot LP	x8 PCIe Gen4 1 Slot LP	x16 PCIe Gen4 2 Slot FHFL	x16 PCIe Gen4 2 Slot FHFL	x16 PCIe Gen4 2 Slot FHFL 1 NVLink Bridge	x16 PCIe Gen4 1 slot FHFL	x16 PCIe Gen4 2 Slot FHFL
Max Power	700W	350W	2x 400W	500W	300W	165W	72W	60W	350W	300W	300W	150W	250W
FP64 TC | FP32 TFLOPS[2]	67 | 67	51 | 51	134 | 134	19.5 | 19.5		10 | 10	NA | 30	NA | 4.5	NA | 91.6	NA | 90	NA | 37	NA | 31	NA | 4x4.5
TF32 TC | FP16 TC TFLOPS[2]	989 | 1979	756 | 1513	1979 | 3958	312 | 624		165 | 330	120 | 242	18 | 36	183 | 366	90 | 181	150 | 300	125 | 250	4x18 | 4x36
FP8 TC | INT8 TC TFLOPS/TOPS[2]	3958 | 3958	3026 | 3026	7916 | 7916	NA | 1248		NA | 661	485 | 485	NA | 72	733 | 734	362 | 724	NA | 600	NA | 500	NA | 4x72
GPU Memory	80GB HBM3 3350 GB/s	80GB HBM2e 2000 GB/s	188GB HBM3 7600 GB/s	80GB HBM2e 2039/1935 GB/s		24GB HBM2 933GB/s	24GB GDDR6 300GB/s	16GB GDDR6 200 GB/s	48GB GDDR6 864 GB/s	48GB GDDR6 864 GB/s	48GB GDDR6 696GB/s	24GB GDDR6 600GB/s	4x 16GB GDDR6 4x 232 GB/s
Multi-Instance GPU (MIG)	Up to 7		UP to 14	Up to 7		Up to 4	-	-	-	-	-	-	-
Media Acceleration	7 JPEG Decoder 7 Video Decoder		14 JPED Decoder 14 Video Decoder	1 JPEG Decoder 5 Video Decoder		1 JPEG Decoder 4 Video Decoder	2 Video Encoder[3] 4 Video Decoder[3] 4 JPEG Decode	1 Video Encoder 2 Video Decoder (+AV1 decode)	3 Video Encoder 3 Video Decoder 4 JPEG Decoder	3 Video Encoder 3 Video Decoder 4 JPEG Decoder	1 Video Encoder 2 Video Decoder (+AV1 decode)		4 Video Encoder 8 Video Decoder (+AV1 decode)
Ray Tracing	-		-	-			Yes	Yes		Yes			
Transformer Engine	Yes		Yes	-			FP8	-	FP8	FP8	-	-	-
DPX Instructions	Yes		Yes	-			-	-			-	-	-
Graphics	For in-situ visualization (no NVIDIA vPC or RTX vWS)			For in-situ visualization (no NVIDIA vPC or RTX vWS)			Better	Good	Top-of-Line	Top-of-Line	Best	Better	Good
vGPU	Yes									Yes			
Hardware Root of Trust	Internal and External			Internal with Option for External					Internal	Internal	Internal with Option for External		
Confidential Computing	Yes			*(1)*		-	-	-	-	-	-	-	-
NVIDIA AI Enterprise	Add-on	Included	Add-on	Add-on						Add-on			

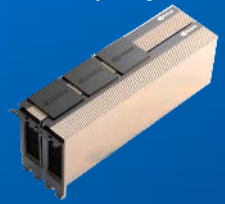
1. *Supported on Azure NVIDIA A100 with reduced performance compared to A100 without Confidential Computing or H100 with Confidential Computing.*
2. *All Tensor Core numbers with sparsity. Without sparsity is ½ the value.*
3. *Includes AV1 in addition to H.265, H.264, VP9, VP8, MPEG4*

DELL Technologies

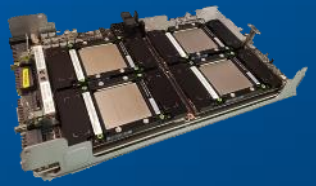
GPU Accelerators

PCIe Adapter

PCIe with
2-way Bridge



4-way SXM / OAM
Baseboard



- Accelerate demanding AI/ML, HPC, data analytics workloads for faster value extraction and collaboration for VDI
- Drive enhanced workload outcomes with greater insights, inferencing and visualization

Brand	Model	Memory	Max Power	Form-Factor	2-way Bridge Capable	Recommended Workloads
PCIe form factor						
AMD	MI210	64 GB HBM2e	300W	DW - FHHL	✓	HPC | AI Training
Intel	Max 1100*	48 GB HBM2e	300W	DW - FHHL	✓	HPC | AI Training
Intel	Flex 140*	12 GB GDDR6	75W	SW - HHHL | FHHL		AI Training
OAM form factor						
Intel	Max 1550*	80 GB HBM2e	600W	OAM with XeLink		AI Training | HPC

Not supported on R760xa

To the edge

Accelerate anywhere

- Dell's 'built-for-the-edge' server portfolio
- Short-depth to fit in field cabinets & racks (<483mm/<19")
- Front-facing I/O to make servicing in tight spaces easier for field engineers
- Shock, vibration, dust, and thermally rated for harsh and unpredictable edge environments (MIL/NEBS)
- Dell ecosystem-enabled with iDRAC

Monolithic

XR7620

- 472mm chassis 2U, 2S Intel® Xeon® Scalable Processors
- Supports 2 x 300W GPUs for AI at Edge
- GPU and CPU-optimized configurations to handle multitude of edge-use cases
- -5C to 55C operating temperature

XR5610

- 463mm chassis 1U, 1S Intel® Xeon® Scalable Processor
- Right-sized for on-site dedicated workloads
- Telco-optimized configuration with time & sync card available
- -5C to 55C operating temperature

Multi-node

XR4000

- 2U multi-node with Intel® Xeon® D
- Dell shortest-depth server at 350mm
- Nano witness-node allows for VM-cluster in single box
- Rackable, stackable, and wall-mountable for ultimate deployment flexibility
- -5C to 55C operating temperature

XR8000

- 2U multi-node with 1S Intel® Xeon® Scalable with optional vRAN boost up to 4 nodes per chassis
- -20C to 65C operating temperature for select configurations
- Telco-optimized for DU and CU RAN deployments
- Extensible to multitude of enterprise edge use cases

Dell Edge Gateway 3200

Rich Storage
- M.2 SSD

Security
- TPM 2.0

Intel Elkhart Lake – Quad Core
- Intel Atom (x6425RE)
- Up to 32GB DDR4 memory

Front accessible I/O and Adaptive uFM module
- RS-232/422/485
- CANbus
- 2 x GbE
- 4 x USB 3
- 2 x DP++
- 2 x Serial COM
- 6-ch DI
- 6-ch DO

Optional Expansion
- CANbus
- PoE out
- 2 x Isolated Serial COM
- 2 x Isolated GbE
- DIO
- PoE in
- Sensors T, P, RH, GYRO

- -20˚C to 60˚C operating temperature
- Compact, Fan-less and Rugged Design
- WiFi-6E, Bluetooth 5.2 and Certified Dual Nano SIM 4G & 5G options
- OEM ready options

EDGE / TELCO

TARGET WORKLOADS

Manufacturing
Simplify and automate data collection at every stage of the production cycle

Telecom
Accelerate innovation and revenue growth with new services

Retail
Personalize customer experience with data insights

Smart Cities
Improve quality of life by increasing the city's efficiency

DELL Technologies

Summary & Conclusion

Accelerate Intelligent Outcomes Everywhere

Align business and IT to a data-first culture

Put any data to work anywhere in any way

Achieve success at any scale as you grow

DELL Technologies

Thank You