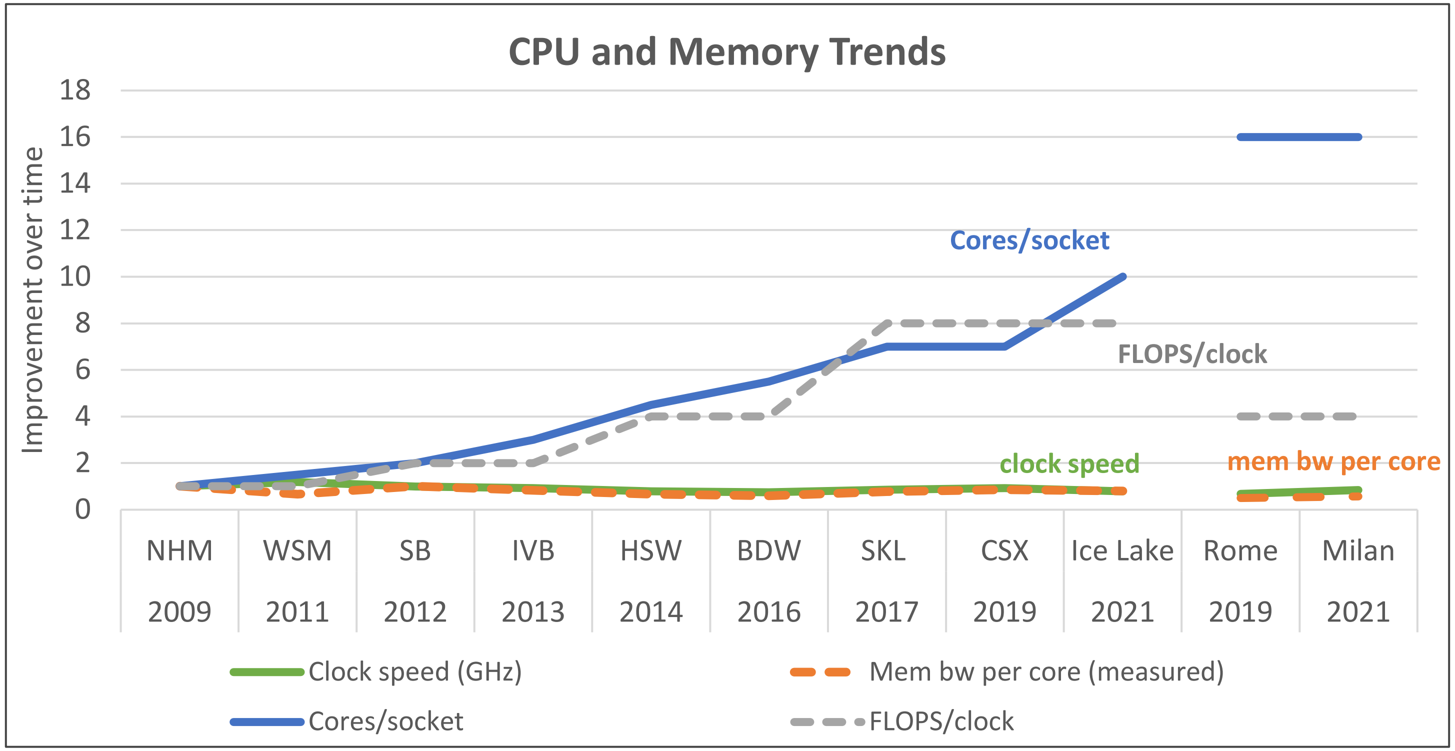
Design Recommendations for HPC DataCenters

Paolo Bianco
EMEA WER HPC and AI Business Dev
paolo.bianco@dell.com

Compute – CPU

- Processor technology
  - Already underway: many many cores but single threaded performance about the same.
  - Chiplet architecture with complex NUMA => more latency challenges, local vs. remote IO.
  - HBM or other fast memory technologies coming with future CPUs.
  - Increasing power requirements with a step function increase. (=> cooling challenges)



Focus in the Future

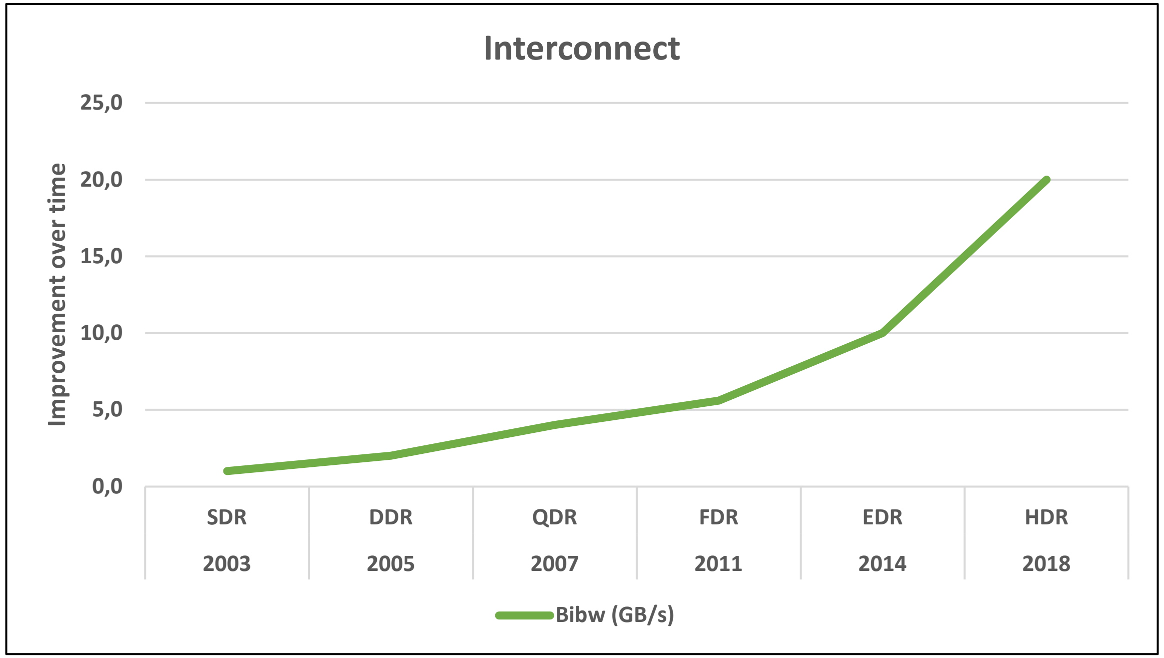
- **Optimize for utilization**

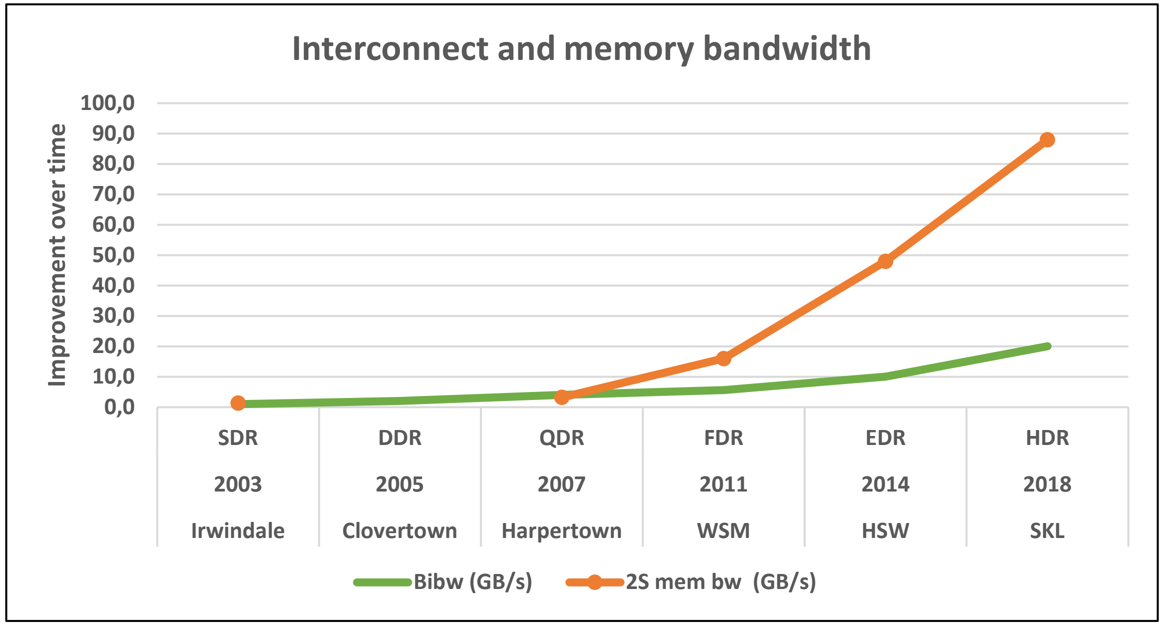
  (Continue to reduce true cost of compute)
- **Improve IO and memory subsystems**

  (Amdahl's Balanced System Law)

Interconnect and Memory

- Memory bandwidth
  - Driven primary by increasing number of processors memory channels.
  - Continues to improve, but at the cost of system footprint and energy.
- Network bandwidth
  - Not keeping pace with system memory bandwidth.
- Results
  - Increased reliance on data localization to improve overall performance

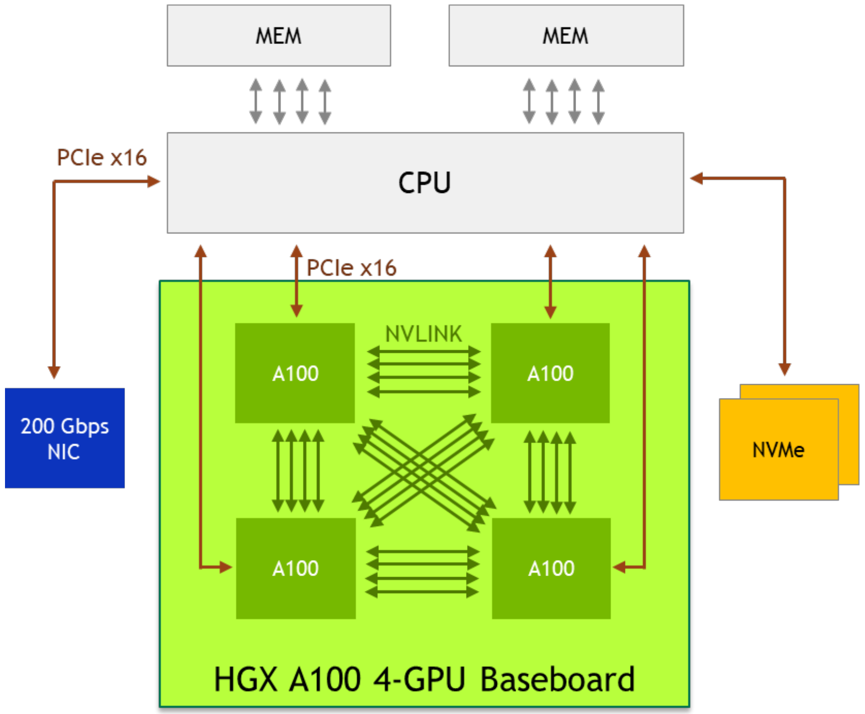




intel. Innovation Built-In DELL Technologies

Compute - GPU

- Many GPU choices with diverse programming models
  - Intel, AMD and Nvidia with oneAPI, ROCm, CUDA
  - Increasing power requirements
  - Many domain specific accelerators for inference, training. GPUs, FPGAs, custom ASICs.
- Cache coherence in the system
  - Reduce copies, access to larger memory capacity for larger datasets.
  - AMD Infinity Fabric CPU to GPU
  - Nvidia Grace CPU – GPU
  - CXL
- Proprietary, new technologies driving custom design requirements
  - Divergent portfolio, custom configurations, cooling and power delivery challenges

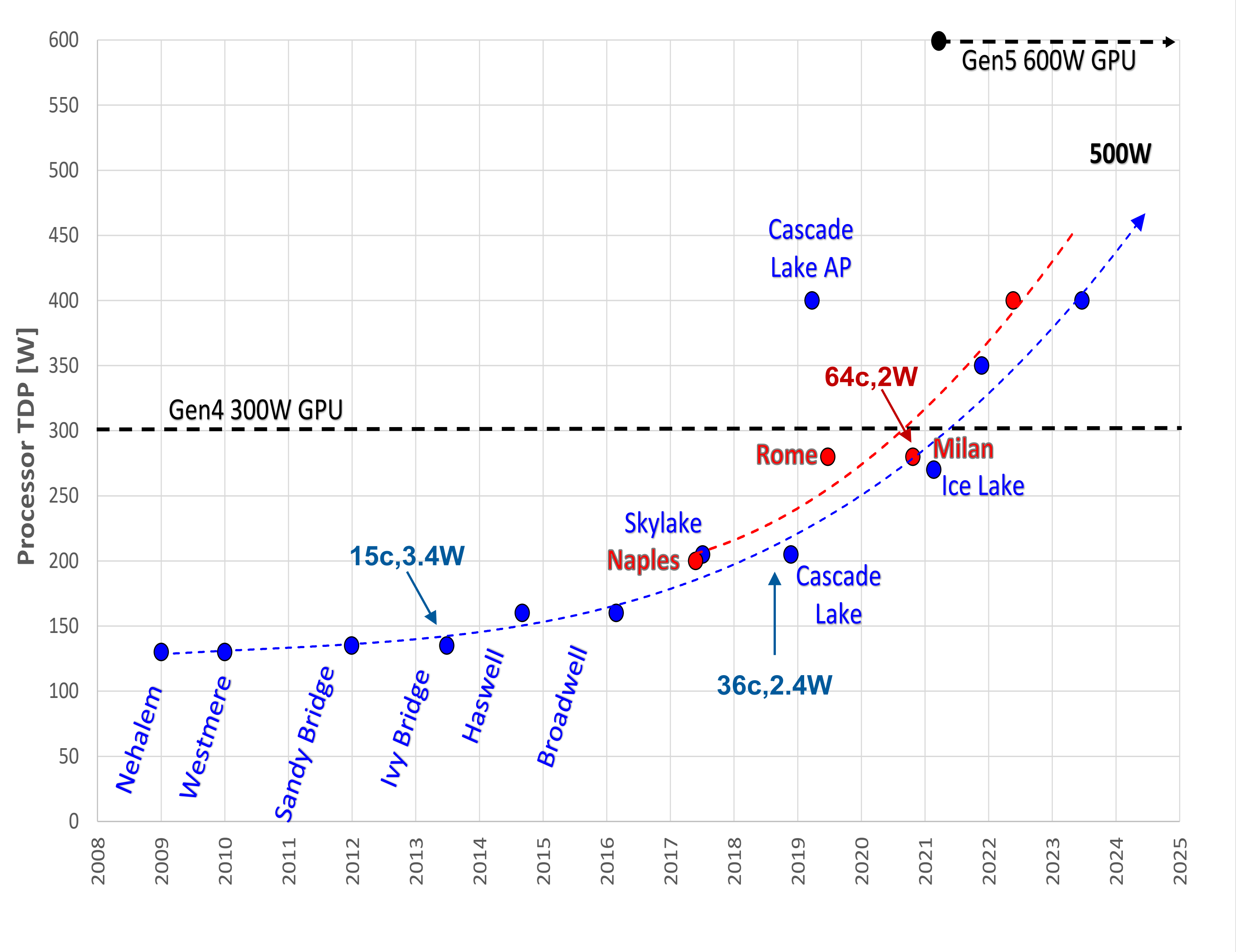


Today’s Challenges

DELL Technologies

Challenge #1: CPU/GPU Thermal Design Power Trends

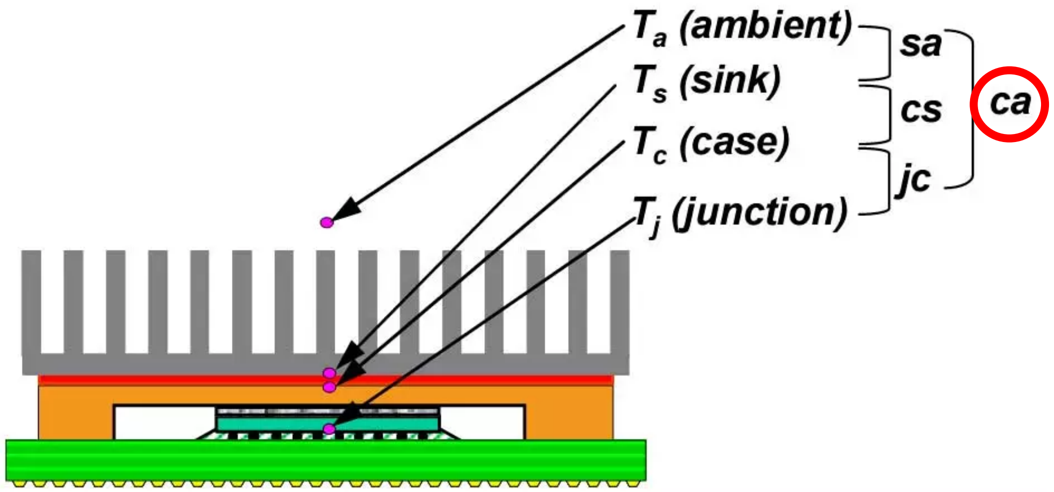
- Competition in the CPU & GPU markets (“Power war”) will continue to drive up power
  - Higher TDPs (with max Tcase reducing!)
  - Higher core number
- Increased Memory count, capacity and speed all adding power
- Accelerated adoption of NVMe, high speed I/O and accelerators also contributing more power
- **Result:** extended air-cooling causing challenges within data center

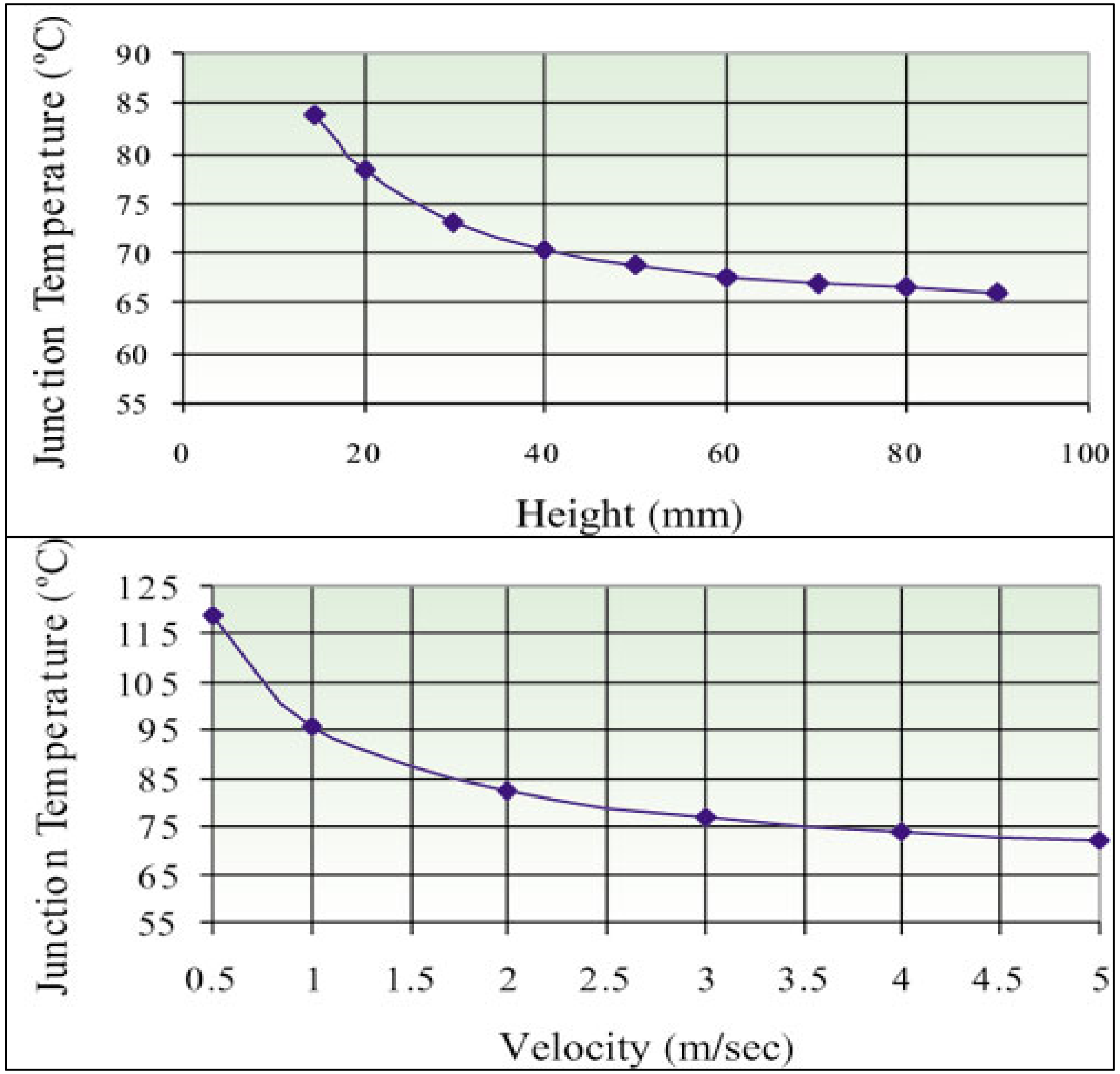


intel. Innovation Built-In DELL Technologies

We can’t increase the size of the heatsinks…

- Physical limitations in heat exchange
  - Speeding up air does not help as well!
- Industry Trend toward reduction of Tcase
- **Result: limitations in max supported chip TDP**

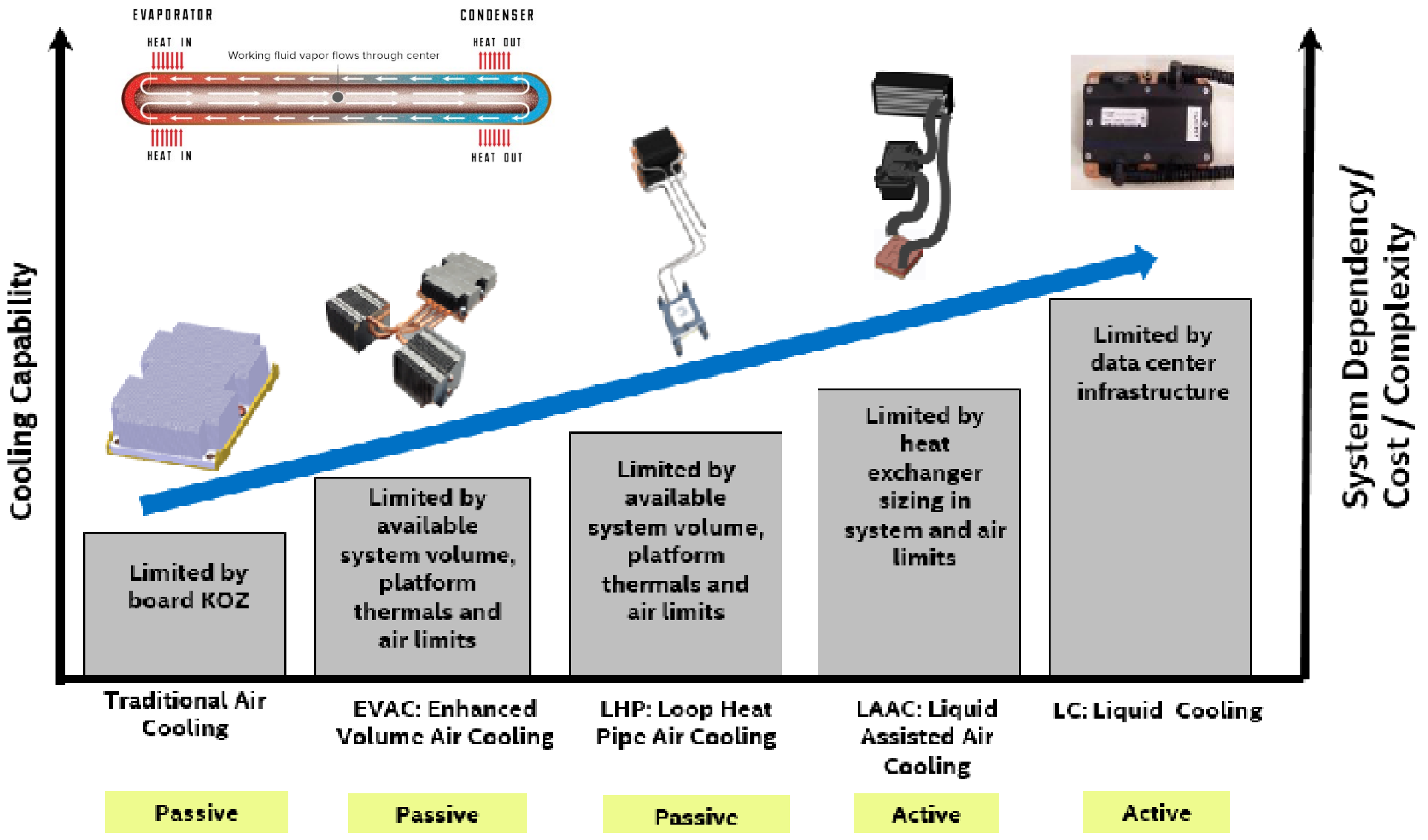




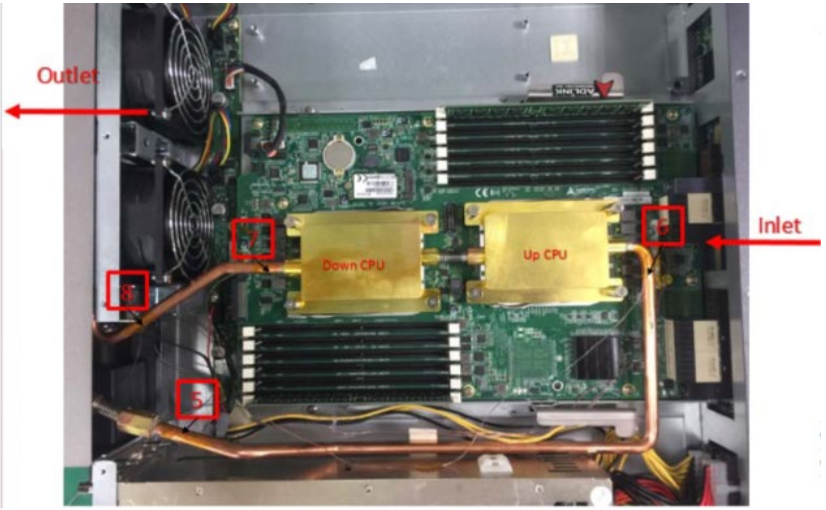
Source: B.Tavassolli - How Much Heat can be Extracted from a heatsink? – Electronics Cooling, 2003

intel. Innovation Built-In DELL Technologies

…but we can spread the heat!



Source: Y Fan, C Winkel et al. - Analytical Design Methodology for Liquid Based Cooling Solution for High TDP CPUs – 17th IEEE Itherm, 2018



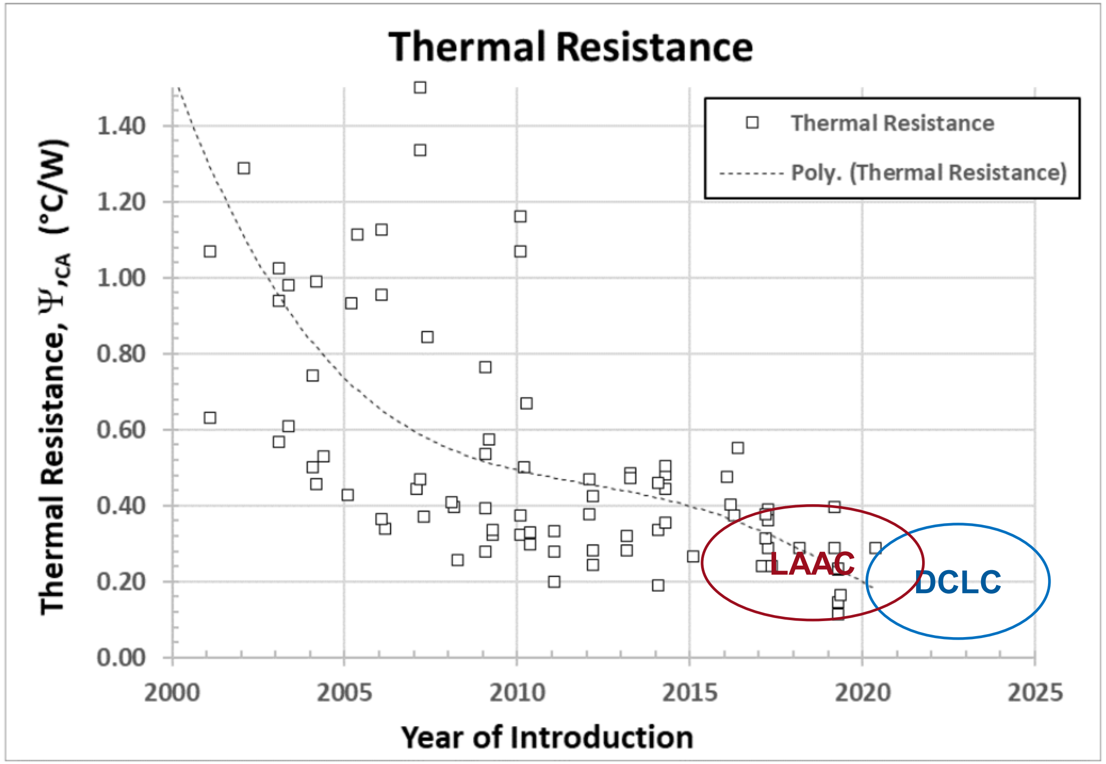
LHP: Loop Heat Pipe (Air Cooling)

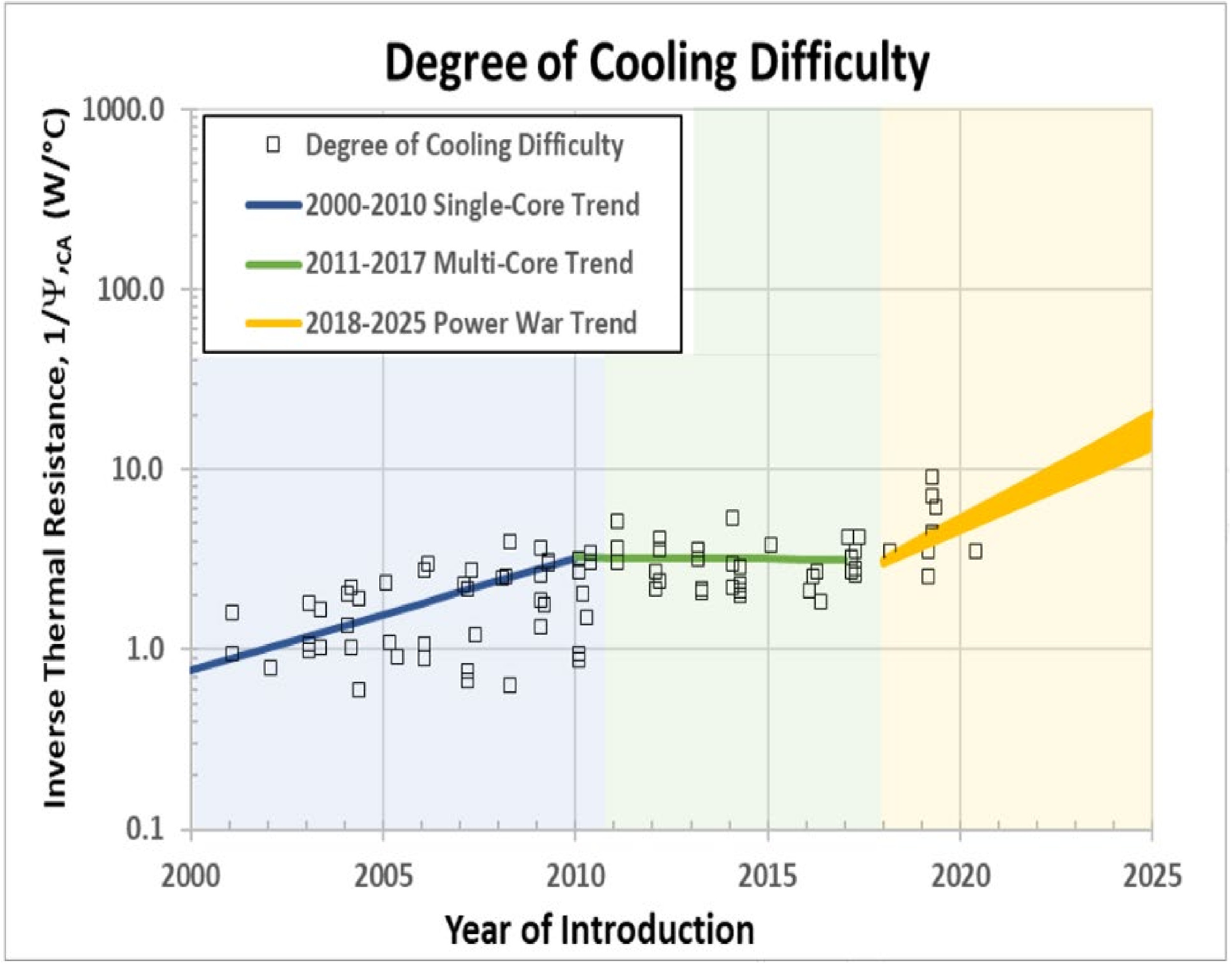
LAAC: Liquid Assisted Air Cooling

Results: Increase in systems size and complexity

A steep change ahead

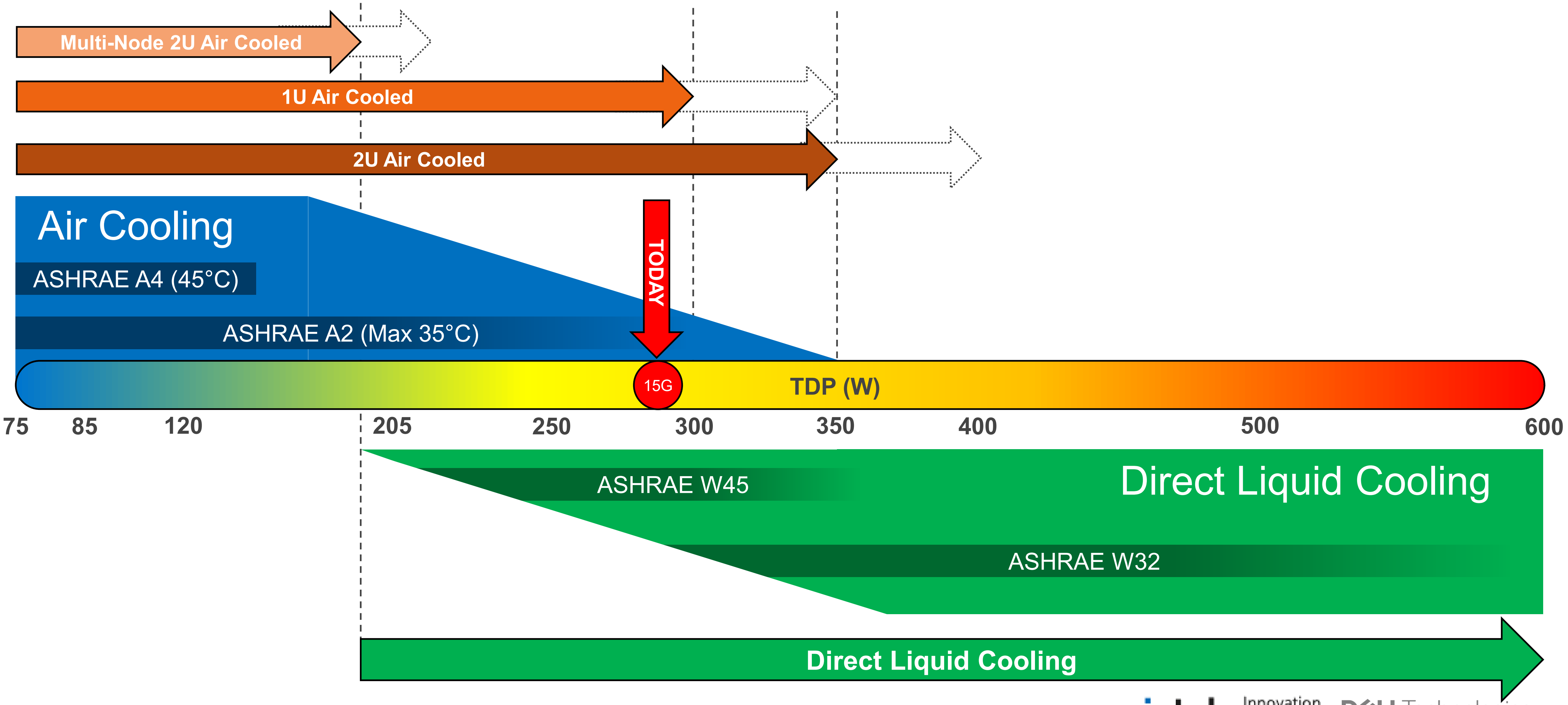
- DoC difficulty ramping up exponentially
- Entering the DCLC Thermal Resistance area





Source : ASHRAE – Emergence and Expansion of Liquid Cooling in Mainstream Data Centers

Air vs. Liquid Cooling Thresholds by Form Factor



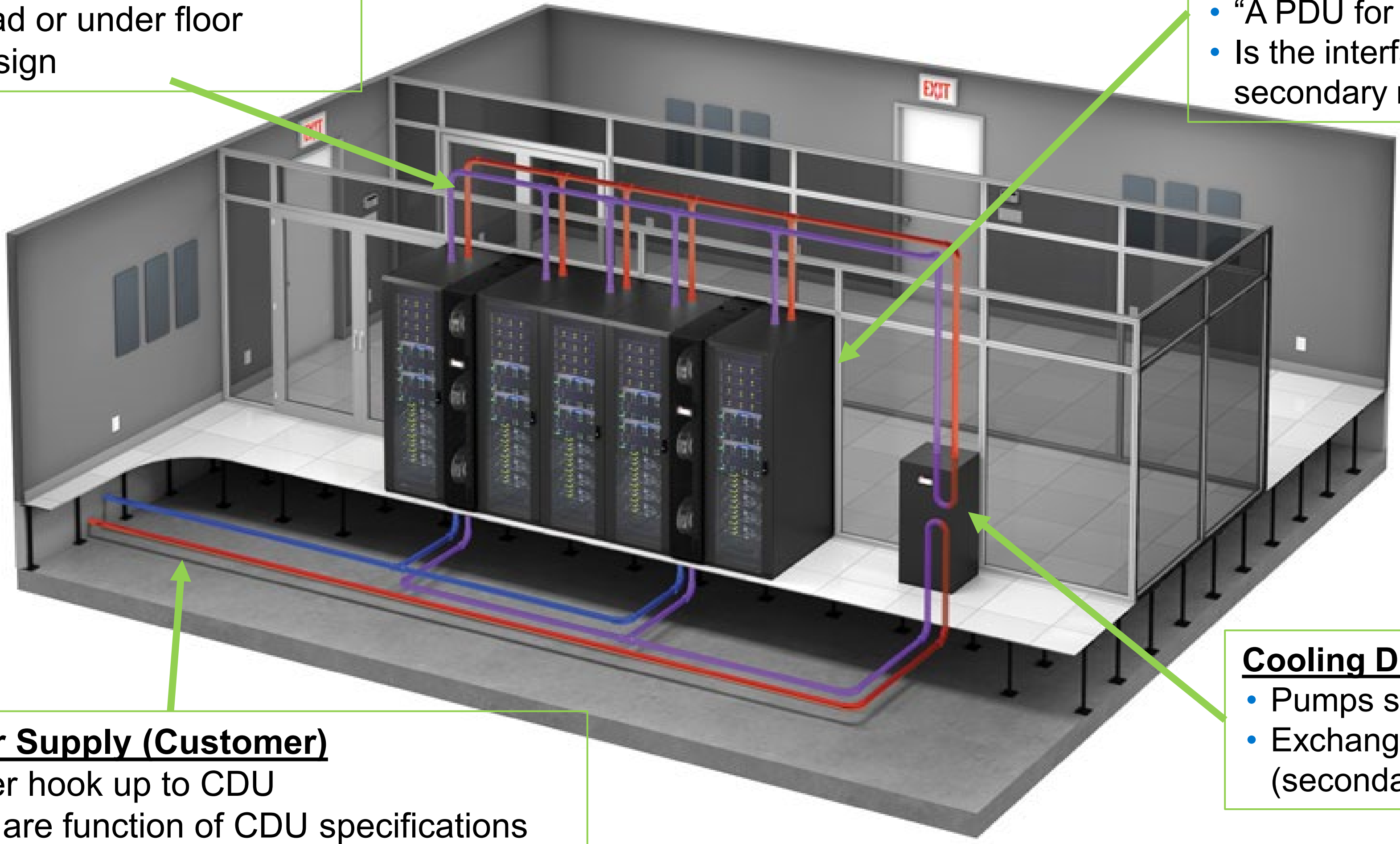
Source : ASHRAE – Emergence and Expansion of Liquid Cooling in Mainstream Data Centers

intel. Innovation Built-In DELL Technologies

Direct Liquid Cooling (DCLC) Ecosystem

<u>Secondary Fluid Network</u>
- Feeds racks from CDU
- Can be overhead or under floor
- Site specific design

<u>Rack Manifold</u>
- Distributes coolant from CDU into server.
- “A PDU for Water”
- Is the interface from the server into the secondary network.



<u>Cooling Distribution Unit, CDU</u>
- Pumps secondary coolant to servers.
- Exchanges heat from server loop (secondary) to building loop (primary)

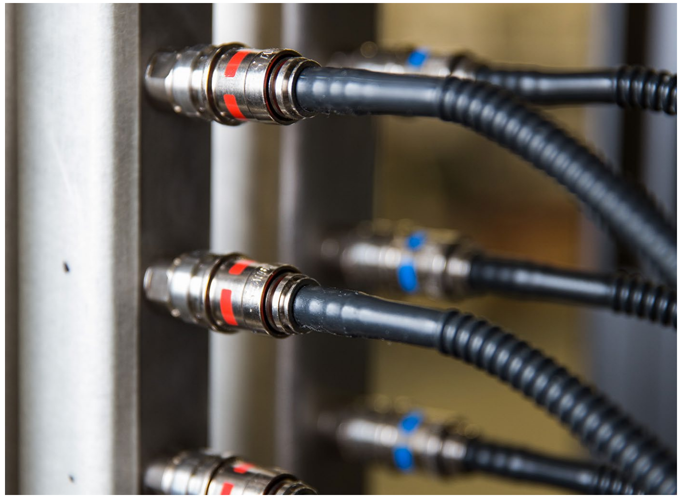
<u>Primary Water Supply (Customer)</u>
- Building water hook up to CDU
- Connections are function of CDU specifications

intel. Innovation Built-In

DELL EMC

DLC: Disadvantages

- A significant portion (about 15-20%) of the total DLC system investment is lost when servers are replaced (coldplate assemblies and tubing are purpose-designed)
- No warranty that different server vendors will adopt the same DLC system in the future (quite all enterprise servers manufacturer are however converging toward Staübli push-fit connectors)
- High water temperatures (usually) require specific cooling systems
- Good Water Quality required on primary circuits
- Using a single cooling system for traditional cooling and DLC may not make sense, as the cooling fluids would usually be at different temperatures.
  - But output water from Rear Door Heat Exchangers could be reused as inlet water for DLC (if DC cooling system can support this)
- Leaks risk
  - Robust leak detection a must
  - Controlling water when it leaks is key

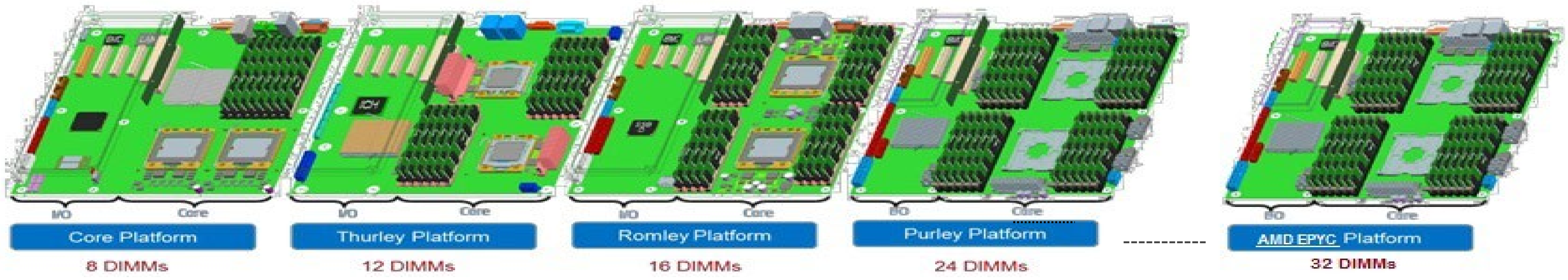


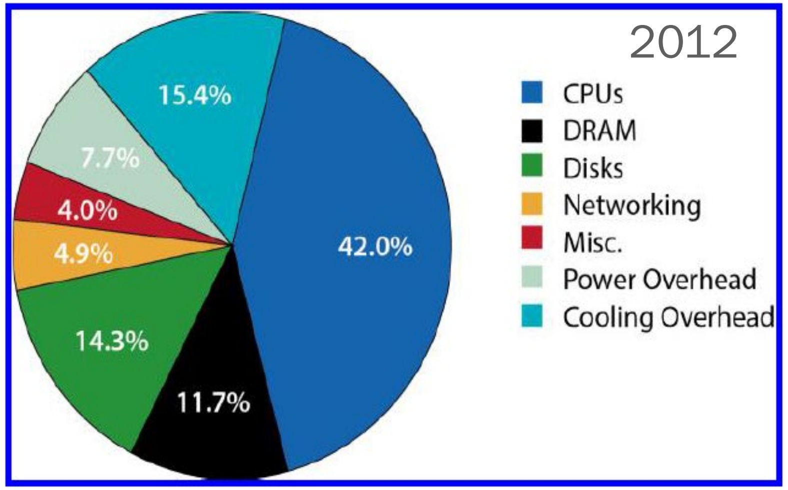
Parameter	Recommended Limits
pH	7 to 9
Corrosion inhibitor	Required
Sulfides	<10 ppm
Sulfate	<100 ppm
Chloride	<50 ppm
Bacteria	<1000 CFU/mL
Total hardness (as $CaCO_3$)	<200 ppm
Residue after evaporation	<500 ppm
Turbidity	<20 NTU (nephelometric)

*Dell recommendation at today is to use Direct Liquid Cooling only **IF** necessary and **WHERE** necessary*

intel. Innovation Built-In DELL Technologies

Node Power Consumption Breakdown





Chung-Ta King - Department of Computer Science - WAREHOUSE-SCALE COMPUTERS, National Tsing Hua University, Taiwan

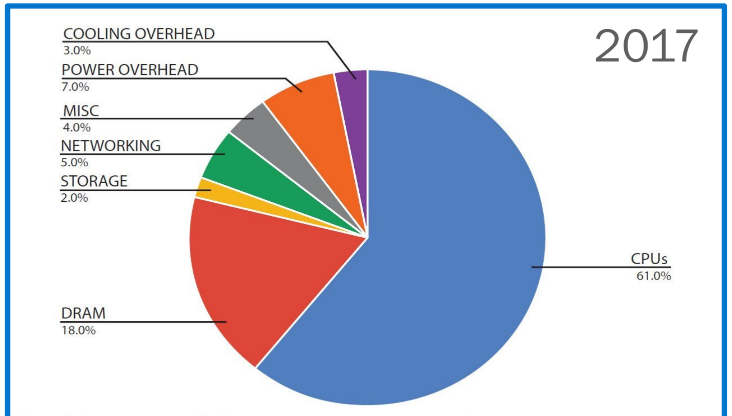
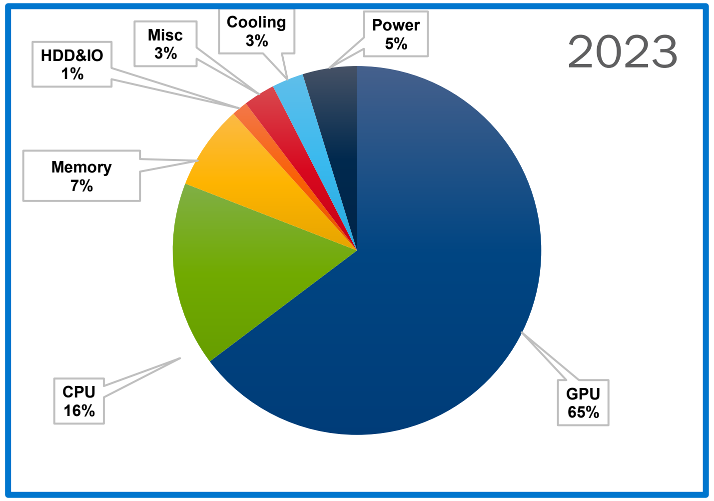


Figure 1.8: Approximate distribution of peak power usage by hardware subsystem in a modern data center using late 2017 generation servers. The figure assumes two-socket x86 servers and 12 DIMMs per server, and an average utilization of 80%.

Source : Barroso, Holzle, Ranghanathan – The Datacenter as a Computer – M&C



Projection considering 2x350W CPU + 4x700W GPU+ 16xDDR5

Intel Innovation Built-In DELL Technologies

No water at the rack: what to do?

- Liquid Assisted Air Cooling "last line of defense" solution (3-5Y horizon)
  - At the expense of system density
  - At the expense of power efficiency
  - Both internal* or external* solutions do do exist
  - Both rack-level or row-level solution exists
- Plan for taller racks, if the case (48U)
- **Bringing water to the rack in the future is key, or the facility will suffer performance limitations!**

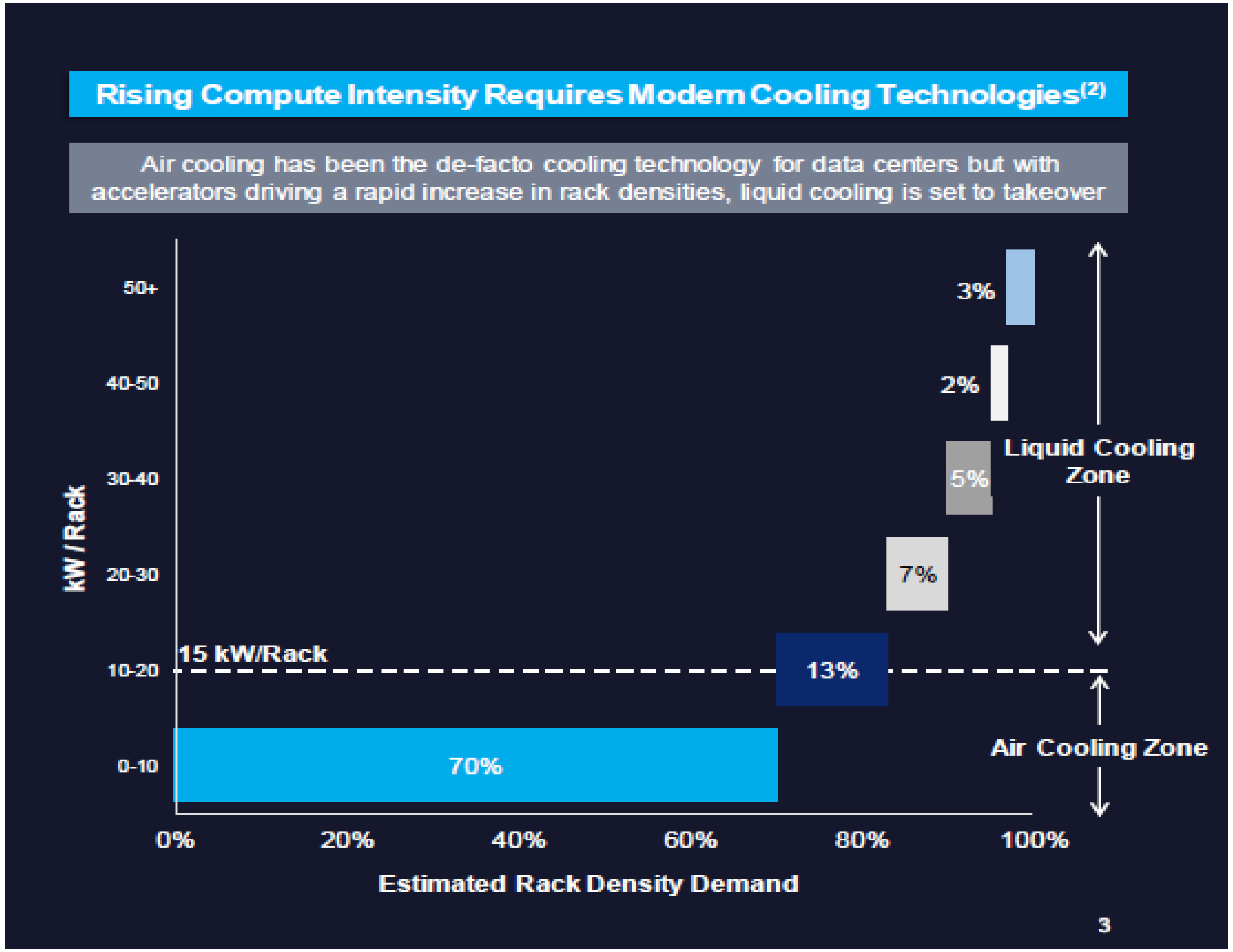
20x60x120cm - 60kW@24°C
1,3kW Power Consumption

7U, 10kW@25°C
750W Power Consumption

* Depending from the system manufacturer

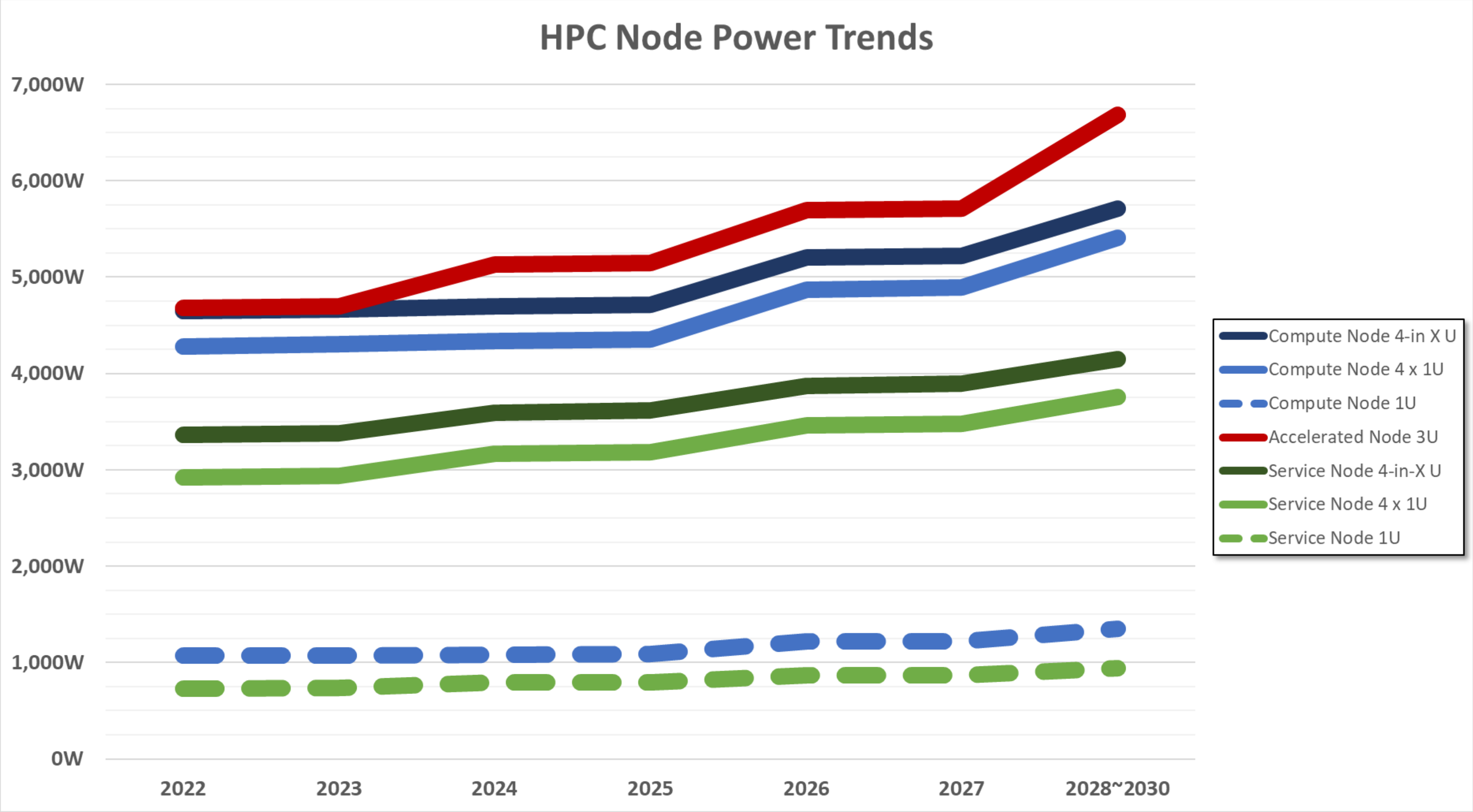
Challenge #2: Rack Power Density Trends

- Common standard density compute Rack loadings are about 12-15kW per Rack
  - only about 25U in use
- Common HPC datacenters loadings are about 30-40kW per Rack
- Path toward exascale requires to cope with 80-90kW per rack.
  - UPS and Power Distribution Architecture needs to be carefully re-engineered.
- *And since Every 1 kiloWatt (kW) of rack power needs 1kW of cooling….*
  - Usual DC Cooling is not going to meet the heat dissipation demand any more
    - Standard cooling not keeping pace with demand
    - Regulatory problems in the DataCenters
    - Vibrations, Noise
    - Air cfm from tiles



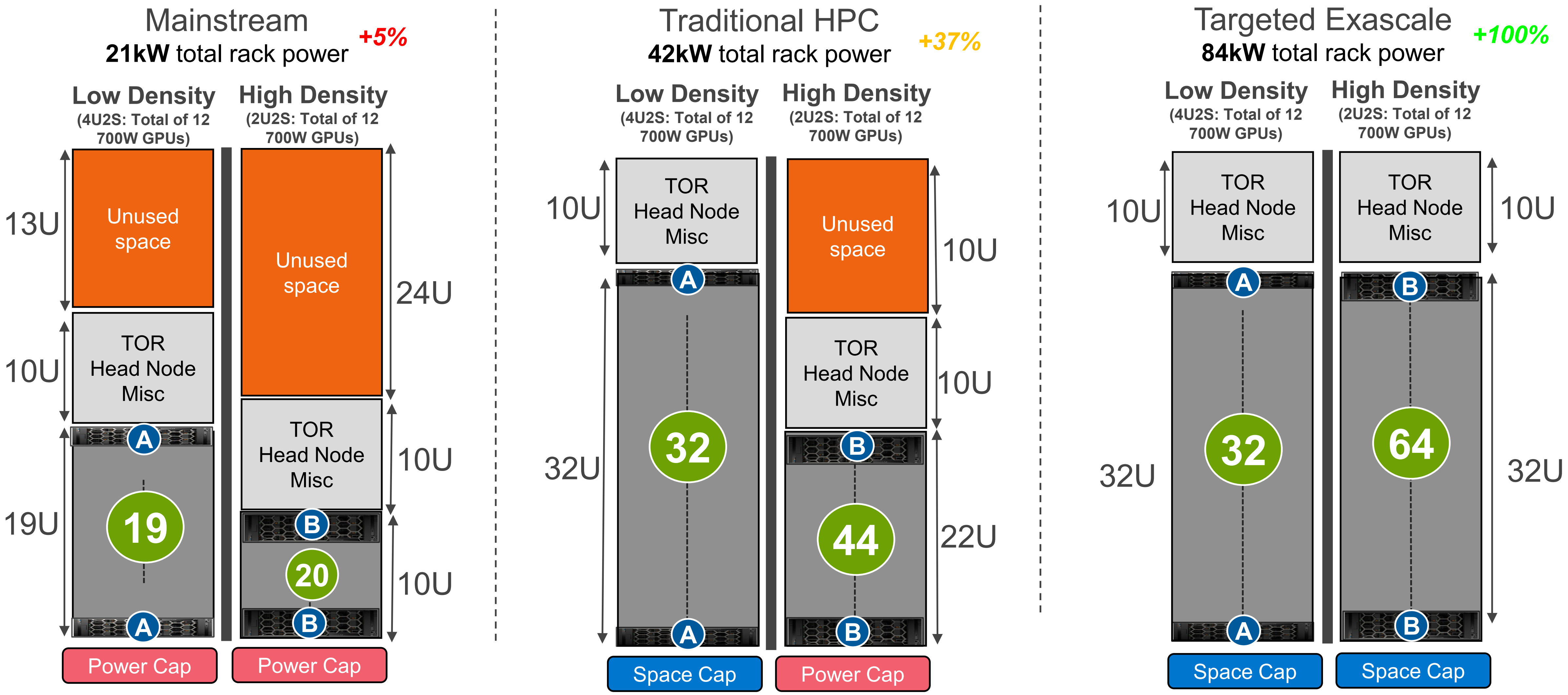
intel. Innovation Built-In DELL Technologies

10 years projection from DT CTO



intel. Innovation Built-In DELL Technologies

CPU Platforms - Density based on 42U Rack



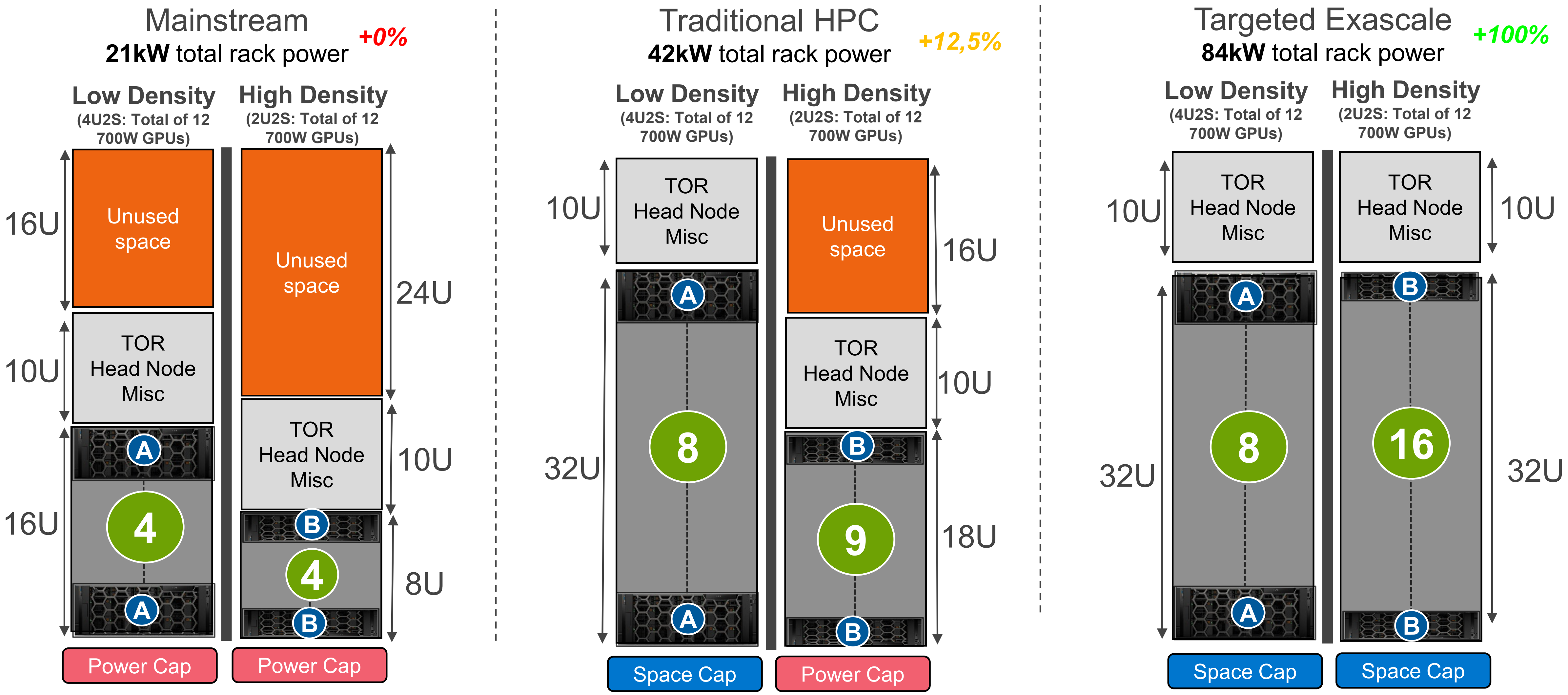
A 1U, 2S, LAAC, Node Power Consumption (sustained)=1 kW

B 2U, 4x2S, DCLC, Chassis Power Consumption (sustained)=3,6kW

Other equipments 2kW

intel Innovation Built-In DELL Technologies

GPU Platforms - Density based on 42U Rack



Ⓐ *4U, 2S+4GPU, HP/LAAC, Node Power Consumption (sustained) = 4.5kW*

Ⓑ *2U, 2S+4GPU, DCLC, Node Power Consumption (sustained) = 4.4kW*

Other equipments 2kW

intel. Innovation Built-In DELL Technologies

Design Recommendations

Design Recommendations for New DataCenters

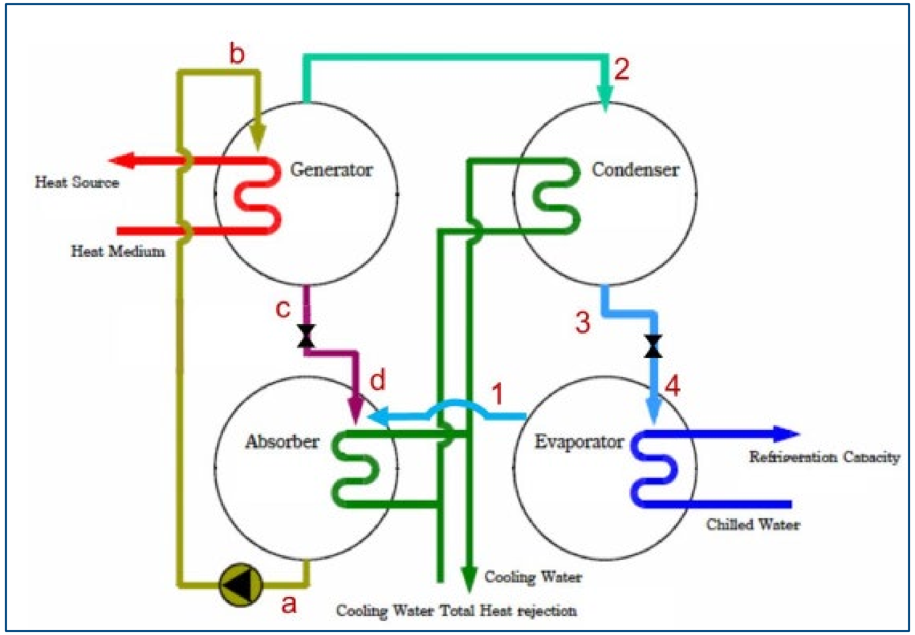
- Plan for 1500mm racks depth
  - Can host proprietary solutions racks or accept multiple PDUs + water cooling manifolds
  - Better if 750mm rack wide
- Plan for 3ph Rack Power Lines
  - At least (2+2) x 32A (42kW) per rack
  - Better if (3+3) x 32A (63kW) or (2+2) x 64A (84kW) per rack
- Plan water to the rack and for 2-temperature water circuits
  - Tempered («high temperature») water (≈ 35-40C) + Cold Water (≈ 12-15C)
  - Power and Network from ceiling, water from floor!
- Plan for good primary water quality
  - A Must for DLC to work properly
- Plan for additional space between racks for CDU (heat exchangers)

Upgrade Recommendations for Existing DataCenters

- Plan for taller racks (ie 48U)
  - May accommodate lower density solutions
  - May accommodate Liquid-to-air exchangers (LAAC)
  - Better if 750mm rack wide
- Plan for 3ph Rack Power Lines
  - At least (2+2)x32A (42kW) per rack
- Plan for water to the rack
  - If possible, tempered («high temperature») water (≈ 35-40C) + Cold Water (≈ 12-15C)
  - Power and Network from ceiling, water from floor!
- Plan for good primary water quality
  - A Must for DLC to work properly
- Plan for additional space between racks for CDU (exchangers)

Key Takeaways

- Chipset Industry focusing on Workload Optimization, memory I/O and data localty
  - No power footprint redution technologies at the horizon
- Direct Liquid Cooling
  - Myth: is the **BEST** cooling method available
  - Reality: is the **ONLY** cooling method available in some cases
    - The ONLY choice is not always the BEST choice
- Dell recommendations:
  - At today, use Direct Liquid Cooling only **IF** necessary and **WHERE** necessary
  - **Plan however for bringing water to the rack,** otherwise you will lose access to an increasing portion of the chipset manufacturers portfolio over time
  - Plan for an upgrade of power lines to the rack, otherwise DataCenter footprints or density issues will arise in the future.
  - Consider ROI of Direct Liquid Cooling more than CPU+GPU before embarking in complex cooling systems acquisitions



- New (old) technology might change the landscape in the future: **Adsorption Chillers**


DELL Technologies
intel
Thank You