Future Computing Architecture

5th and 6th Lessons

Marco Briscolini, PhD

marco.briscolini@gmail.com

Cell: 3357693820

04/17/2024

Piano del Corso – 16 ore in 8 moduli

Descrizione generale delle architetture HPC e AI e loro componenti di base
- Le previsioni di mercato AI&HPC nel mondo
- Componenti principali: parte computazionale, rete di interconnessione, sottosistema storage
- Concetti di metrica delle varie componenti (misurazione della capacita' computazionale, trasmissione dati, lettura/scrittura dati)
- Metriche riconosciute a livello mondiale (Top500, Green500, IO500)
- Concetti introduttivi sull'analisi della complessita' computazionale di un ambito applicativo

Architetture di calcolo e loro evoluzione
- Architetture omogenee e accelerate
- Concetti generali sui microprocessori (CPU)
- Concetti generali sugli acceleratori (Graphical Processor Unit)
- Integrazione CPU-GPU e trasmissione dati

Reti a alte prestazioni per architetture HPC e AI e loro evoluzione
- Reti con protocollo Infiniband e alcune topologie correlate
- Reti di tipo Ethernet a alte prestazioni
- Protocolli RDMA e RoCE

Sottosistemi storage a alte prestazioni e loro evoluzione
- Concetti generali sulla gerarchia dei sottosistemi storage
- Sistemi a disco magnetico e a stato solido
- Connessione di sistemi storage su SAN, Infiniband, Ethernet, nVME over Fabric, e altro

Architetture storage a alte prestazioni
- Architetture di sottositemi storage
- Filesystem paralleli per lettura/scrittura a alte prestazioni

06

Problematiche di efficientamento energetico per sistemi HPC a grande scala (architetture pre e exascale)
- Il concetto di PUE e di efficienza energetica a parita' di potenza computazionale
- Come le varie architetture si caratterizzano in termini di "Potenza di Calcolo"/Watt
- Utilizzo di tecniche di gestione del carico di lavoro per ottimizzare l'efficienza energetica
- Soluzioni di raffreddamento a aria, a acqua diretta e immersivo
- Concetti generali sul disegno e la realizzazione di Data Center efficienti

07

Accenni sulle architetture innovative in ambito AI&HPC
- Architetture AI scalabili
- Interconnessione tra sistemi AI
- AI/HPC/Q-C architettura integrata per carichi computazionali complessi

08

Accenni al disegno e alla progettazione di un'architettura HPC
- Definizione di specifiche di progetto
- Valutazione preliminare dell'architettura ottimale
- Disegno di massima dell'architettura

- Concetto di rispondenza e verifica alle specifiche di progetto

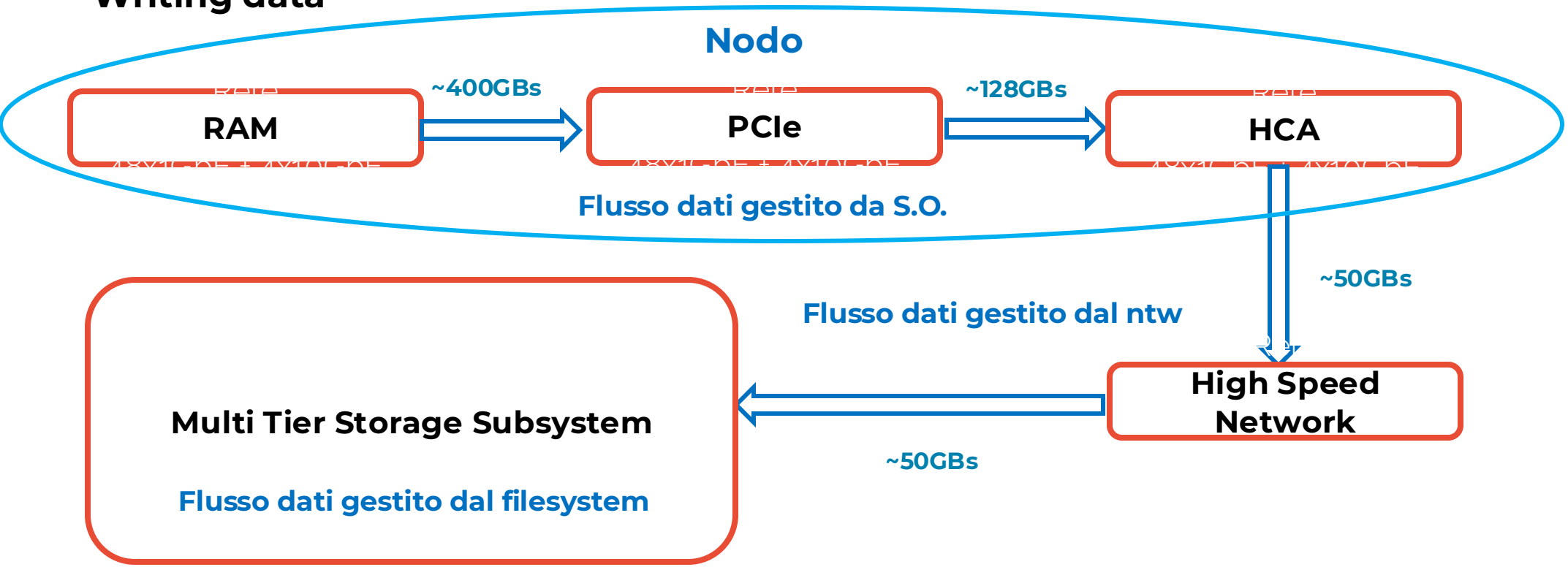
Piano del Corso – Lesson 5

Architetture storage a alte prestazioni

- Architetture di sottositemi storage
- Classificazione logica delle architetture storage
- Accenni su filesystem paralleli per lettura/scrittura a alte prestazioni
- Accenni su architetture storage a multi livello

Il flusso dati: dal nodo al sottosistema storage e viceversa

Writing data



I dati di transfer rate sono teorici, nella realta' i valori misurati sono circa 80% dei valori teorici o anche inferiori

Reading data: flusso inverso

Architetture sottosistemi storage e loro caratteristiche principali

Dischi a stato solido SSD e nVME

- **NVMe concepts**
  NVMe (non-volatile memory express) is a host controller interface and storage protocol created to accelerate the transfer of data between enterprise and client systems and solid-state drives (SSD) over a computer's high-speed Peripheral Component Interconnect Express (PCIe) bus

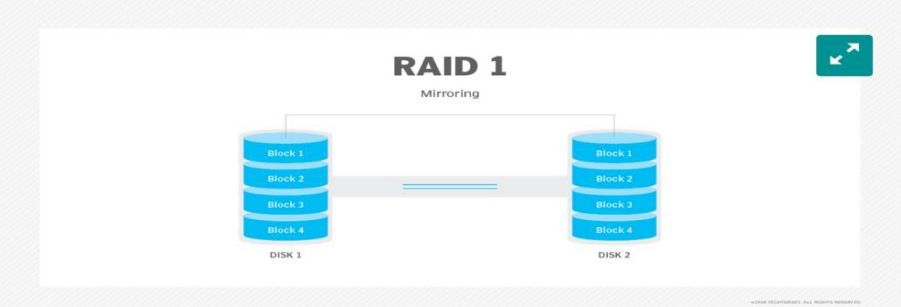
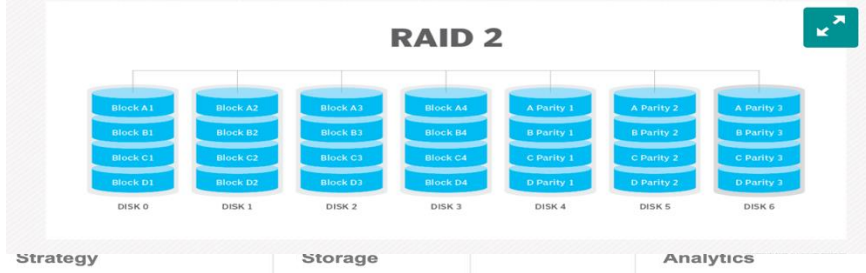
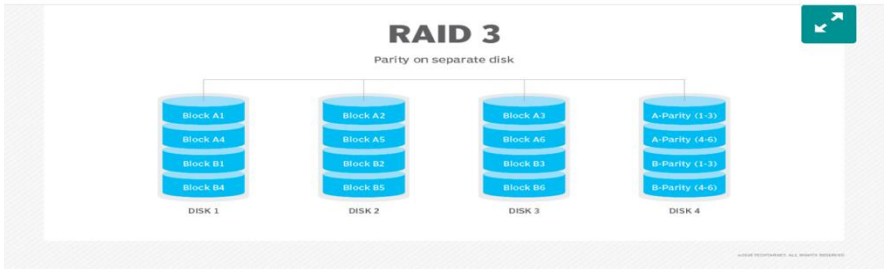
Dischi magnetici: Hard Disk Drive (HDD)

Nastri magnetici: Tape

Architetture sottosistemi storage e loro caratteristiche principali

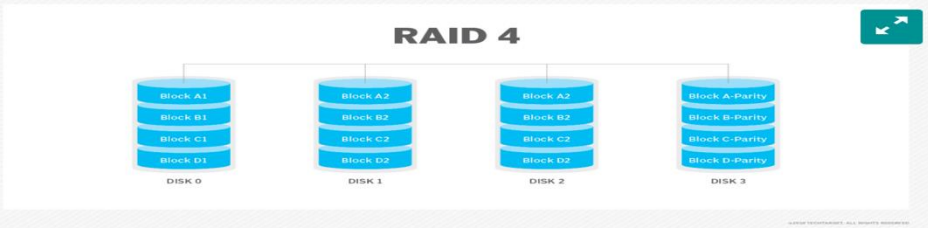
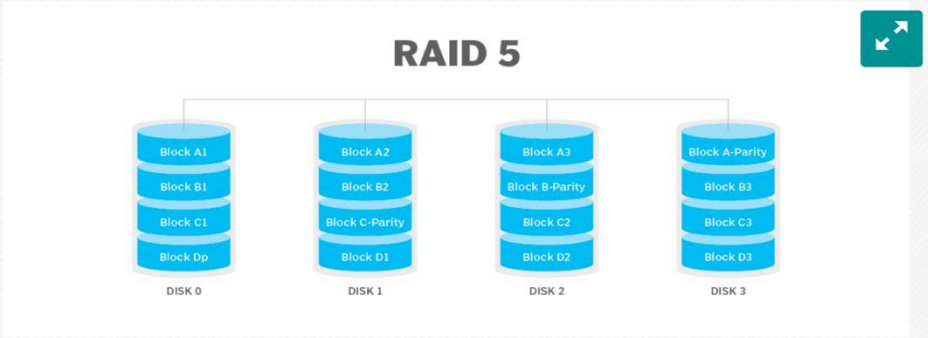
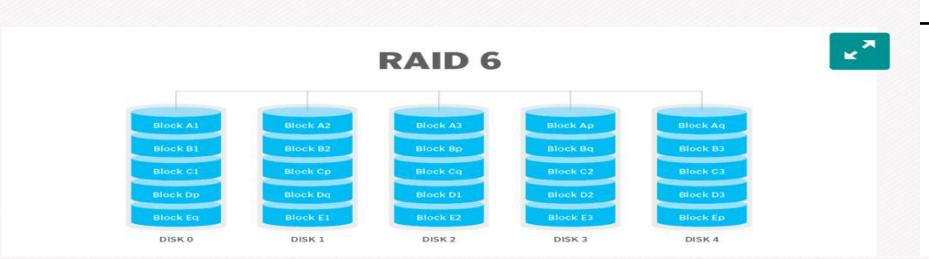
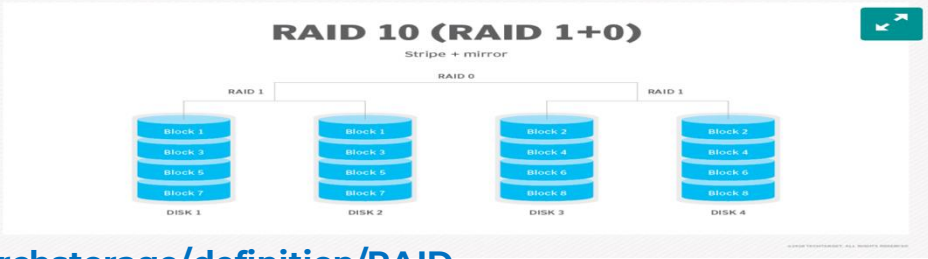
Descr/Tipo	Flash	HDD	Tape	Nota
Capacita' dati per Unita	~10TB	~20TB	~50TB	Le capacita' variano anno/anno
Tipologia di accesso al dato	random	sequenziale	sequenziale	Maggior utilizzo
IOPS e BW	~100000 ~10GBs	~100 ~1GBs	n.a. ~1GBs	Valori molto differenti a seconda del media
Watt per unita'	~10W	~10W	~1W	Valore medio considerando anche lo stato idle
Costo/GB	~0,1USD	~0,03USD	~0,003USD	Costo/GB si riduce anno/anno
Unita' per sottosistema	~decine	~centinaia	~migliaia	Architetture capacitive
Capacita' tipo per sottosistema	~PB	~10PB	~50PB	Valori indicativi

Protezione RAID – Redundant Array of Independent Disk

Descr/Tipo	Schema	Nota
RAID 0		This configuration has striping but no data redundancy. It offers the best performance, but it does not provide fault tolerance.
RAID 1		Also known as *disk mirroring*, this configuration consists of at least two drives that duplicate the storage of data. There is no striping. Read performance is improved since either disk can be read at the same time. Write performance is the same as that of single-disk storage. RAID 1 is an effective tool for disaster recovery.
RAID 2		This configuration uses striping across drives, with some drives storing error-correction code (ECC) information. RAID 2 also uses dedicated Hamming code parity, a linear form of ECC. RAID 2 has no advantage over RAID 3 and is no longer used.
RAID 3		uses striping and dedicates one drive to store parity information. The embedded ECC information is used to detect errors. Data recovery is accomplished by calculating the exclusive information recorded on the other drives. Because an I/O operation addresses all the drives at the same time, RAID 3 cannot overlap I/O. For this reason, RAID 3 is best for single-user systems with long-record applications.

https://www.techtarget.com/searchstorage/definition/RAID

Protezione RAID – Redundant Array of Independent Disk

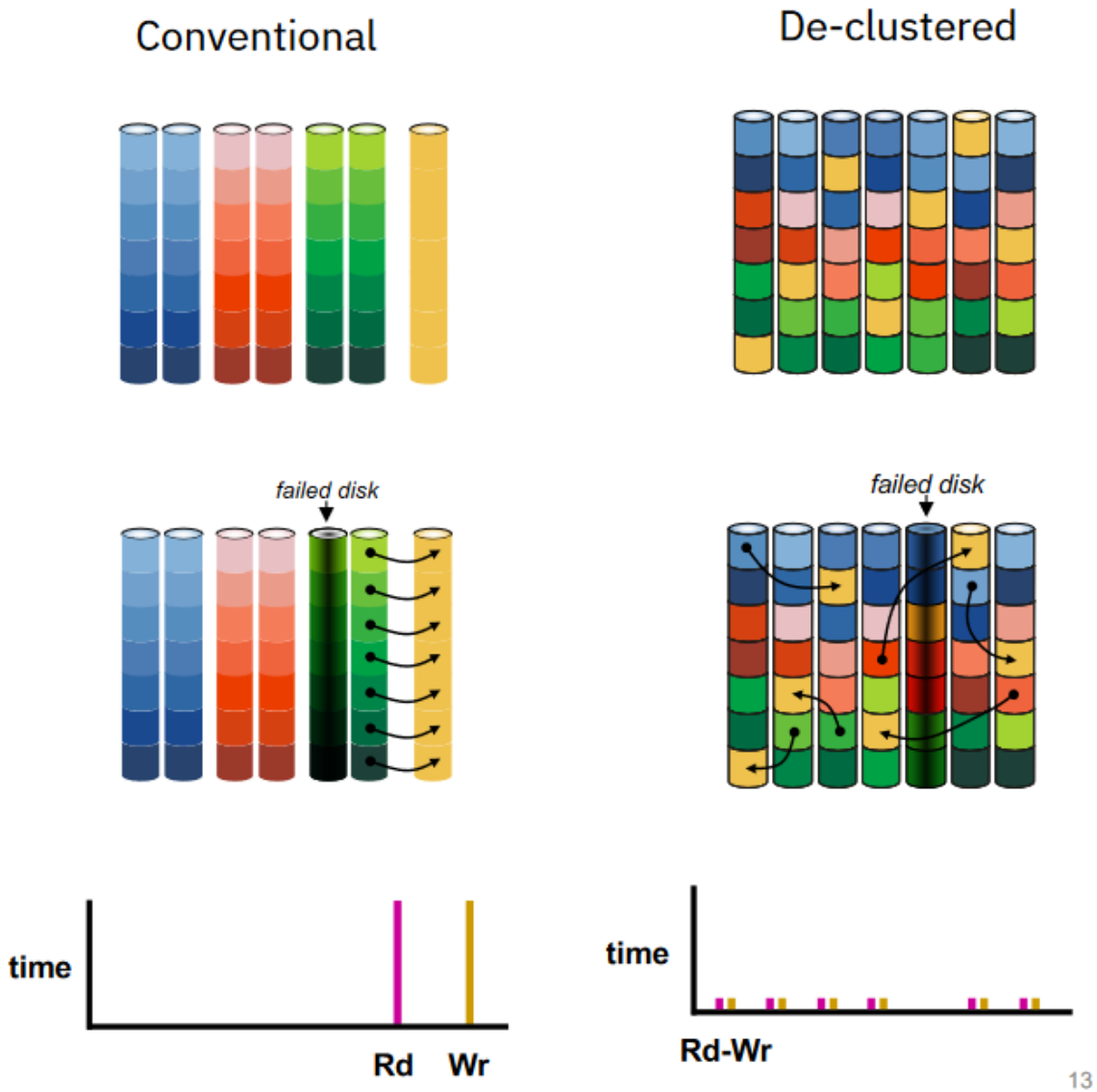
Descr/Tipo	Schema	Nota
RAID 4		uses large stripes, which means a user can read records from any single drive. Overlapped I/O can then be used for read operations. Because all write operations are required to update the parity disk drive, no I/O overlapping is possible.
RAID 5		is based on parity block-level striping. The parity information is striped across each drive, enabling the array to function, even if one drive were to fail. The array's architecture enables read/write operations to span multiple drives. This results in performance better than that of a single drive but not as high as a RAID 0 array. RAID 5 requires at least three drives, but it is often recommended to use at least five for performance reasons. RAID 5 arrays are generally considered to be a poor choice for use on write-intensive systems because of the performance impact associated with writing parity data. When a disk fails, it can take a long time to rebuild a RAID 5 array.
RAID 6		is similar to RAID 5, but it includes a second parity scheme distributed across the drives in the array. The use of additional parity enables the array to continue functioning, even if two drives fail simultaneously. However, this extra protection comes at a cost. RAID 6 arrays often have slower write performance than RAID 5 arrays.
RAID 10 (1+0)		which combines RAID 1 and RAID 0, offers higher performance than RAID 1 but at a much higher cost. In RAID 10, the data is mirrored, and the mirrors are striped.

https://www.techtarget.com/searchstorage/definition/RAID

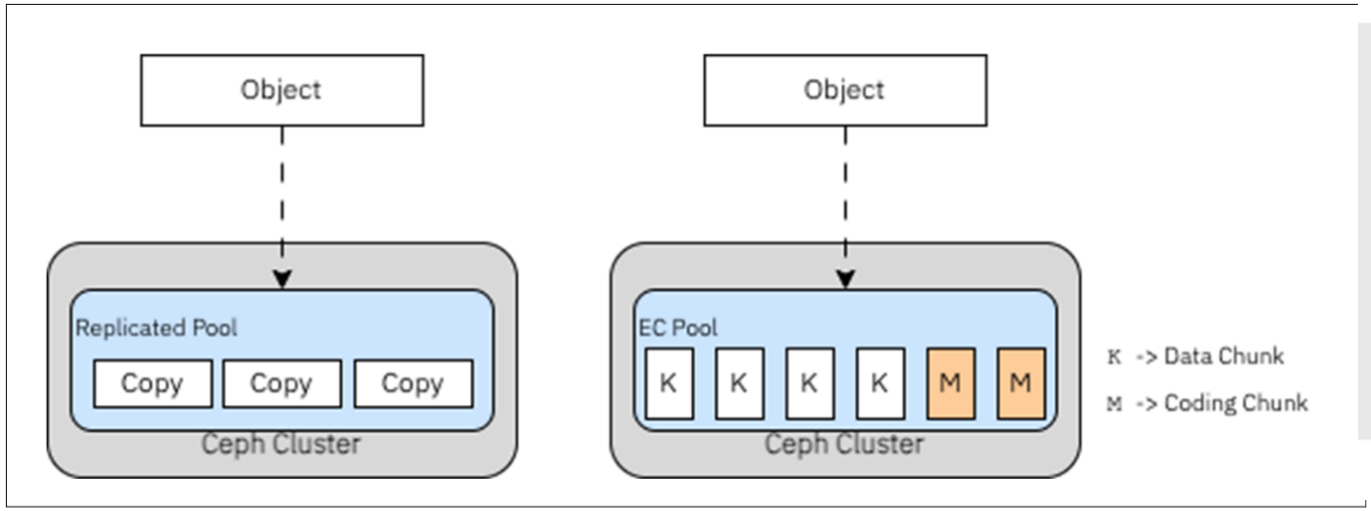
IBM Storage Scale Server – A Brief Detour on Storage Scale RAID

Storage Scale RAID is a *software* implementation of "declustered RAID" that is unique to the ESS/SSS

- A Standard Scale deployment can generally leave the drive media management (integrity, performance, block placement, etc) to the server or storage array
- Since the drives exist in the same enclosure, the Scale software must assume responsibility
- Storage Scale RAID uses Declustered RAID and Erasure Coding to provide:
  - Extremely fast rebuild times with minimal impact on performance
  - Very strong data integrity checks
  - Error detection codes enable detecting track errors and dropped writes
  - Consistent performance from 0 – 99% utilization or 1 to many jobs in parallel



Architettura Ceph – un'architettura cluster per la gestione del dato protetto da errori HW e SW



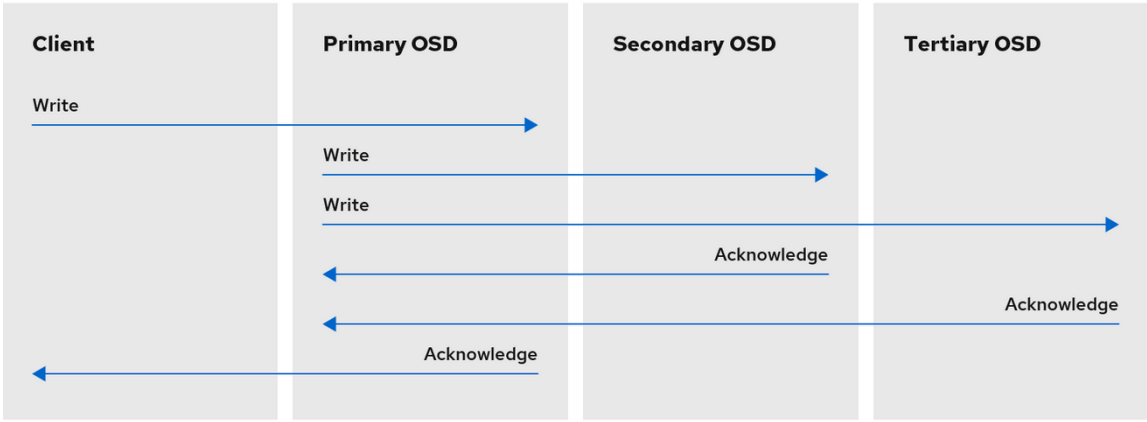


Figure 2-4 Replicated data protection versus erasure coding data protection

The benefits of the replicated model are:

- Very high durability with 3 copies.
- Quicker recovery.
- Performance optimized.

The benefits of the erasure coding model are:

- Cost-effective durability with the multiple coding chunks.
- More expensive recovery.
- Capacity optimized.

Ceph OSD stands for Ceph Object Storage Daemon. It's a program that stores data on a local file system and provides access to it over a network. OSDs are the core of the Ceph storage platform

The classical way to recover from failures in storage systems was to use replication. However, replication incurs significant overhead in terms of wasted bytes. Therefore, increasingly large storage systems, such as those used in data centers use erasure-coded storage. The most common form of erasure coding used in storage systems is Reed-Solomon (RS) code, an advanced mathematics formula used to enable regeneration of missing data from pieces of known data, called parity blocks. In a (k, m) RS code, a given set of k data blocks, called "chunks", are encoded into (k + m) chunks. The total set of chunks comprises a *stripe*. The coding is done such that as long as at least k out of (k + m) chunks are available, one can recover the entire data. This means a (k, m) RS-encoded storage can tolerate up to m failures.

https://www.redbooks.ibm.com/redpieces/pdfs/redp5721.pdf

Erasure code – un algoritmo che protegge i dati in caso di rotture minimizzando l'overhead

The classical way to recover from failures in storage systems was to use replication. However, replication incurs significant overhead in terms of wasted bytes. Therefore, increasingly large storage systems, such as those used in data centers use erasure-coded storage. The most common form of erasure coding used in storage systems is Reed-Solomon (RS) code, an advanced mathematics formula used to enable regeneration of missing data from pieces of known data, called parity blocks. In a (*k*, *m*) RS code, a given set of *k* data blocks, called "chunks", are encoded into **(*k* + *m*) chunks**. The total set of chunks comprises a *stripe*. The coding is done such that as long as at least ***k* out of (*k* + *m*) chunks are available**, one can recover the entire data. This means a (*k*, *m*) RS-encoded storage can tolerate up to *m* failures.

Example: In RS (10, 4) code, which is used in Facebook for their HDFS,[16] 10 MB of user data is divided into ten 1MB blocks. Then, four additional 1 MB parity blocks are created to provide redundancy. This can tolerate up to 4 concurrent failures. The storage overhead here is 14/10 = 1.4X.

In the case of a fully replicated system, the 10 MB of user data will have to be replicated 4 times to tolerate up to 4 concurrent failures. The storage overhead in that case will be 50/10 = 5 times.

This gives an idea of the lower storage overhead of erasure-coded storage compared to full replication and thus the attraction in today's storage systems.

Initially, erasure codes were used to reduce the cost of storing "cold" (rarely-accessed) data efficiently; but erasure codes can also be used to improve performance serving "hot" (more-frequently-accessed) data.[12]

RAID N+M divides data blocks across N+M drives, and can recover all the data when any M drives fail.[1] In particular, RAID 7.3 refers to triple-parity RAID, and can recover all the data when any 3 drives fail.[17]

https://en.wikipedia.org/wiki/Erasure_code **https://min.io/product/erasure-code-calculator**

Concetti generali delle architetture storage scalabili: classificazione fisica

Share nothing storage architecture
Ogni singolo nodo componente il cluster ha uno o piu' dischi interni che non sono condivisibili con gli altri nodi dove risiedono I dati dell'applicazione oltre a altri dati necessari all'utilizzazione delle risorse (home directory, OS, etc). Si possono rendere I dischi interni utilizzabili in toto o in parte anche da altri nodi tramite soluzioni SW che consentono la condivisione delle risorse tramite la rete di interconnessione (share anything)

Scale-out storage architecture
Architetture storage granulari in forma di cluster il cui building-block unitario
puo' essere anche un server con dischi interni (share anything)

Cluster di sottosistemi storage
Architetture storage che raggruppa tramite un ambiente SW piu' sottosistemi storage generalmente simili/identici per comporre un sistema di maggiori dimensioni (share anything)

Scale-up storage architecture

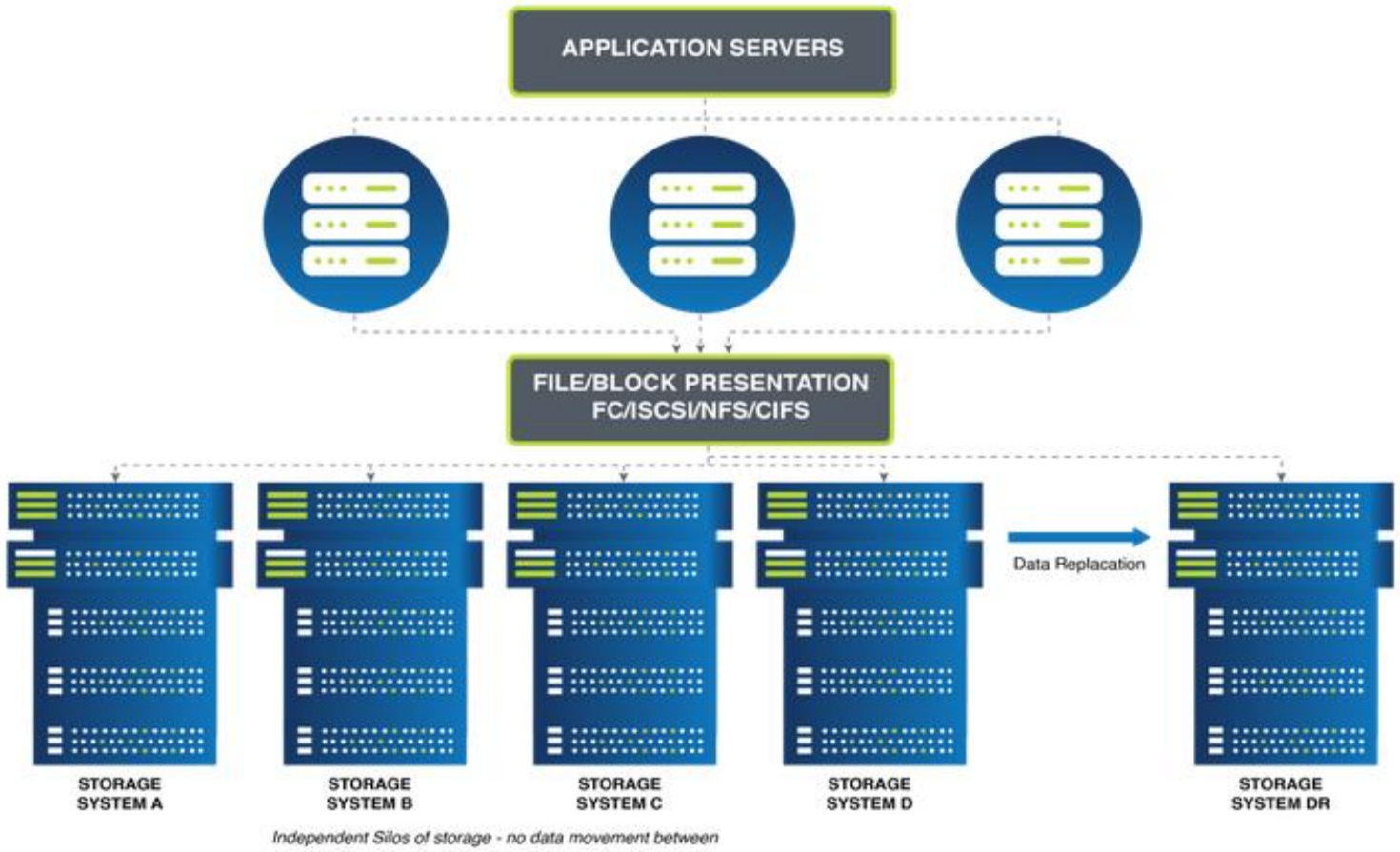


Figure 2 – Modular/Scale-up Storage Silos

https://cloudian.com/blog/scale-up-vs-scale-out-storage/

Scale-out storage architecture

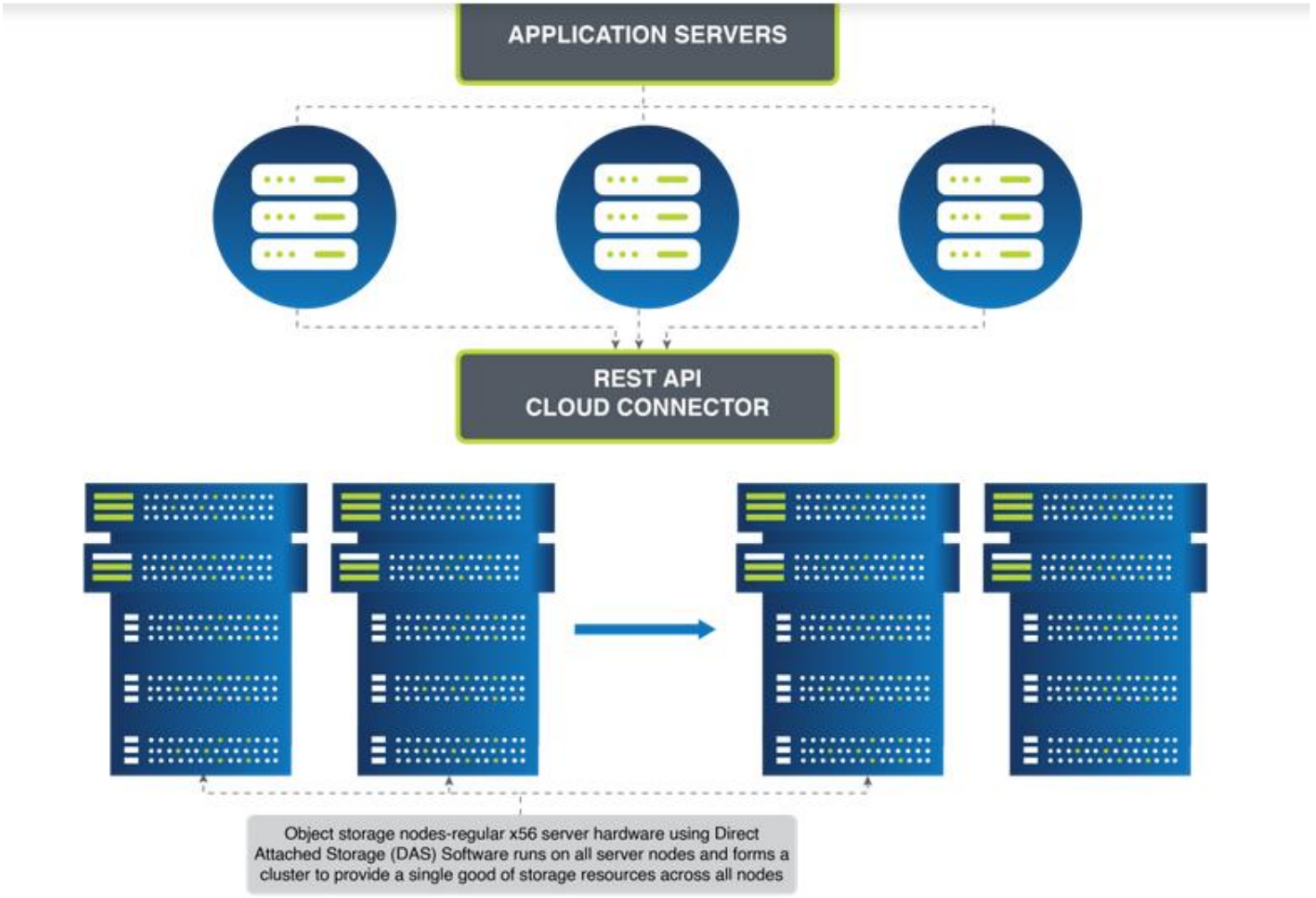


Figure 3 – Object/Scale-out Storage Architecture

https://cloudian.com/blog/scale-up-vs-scale-out-storage/

Concetti generali delle architetture storage scalabili: classificazione logica

Block storage

Divide i dati in blocchi identici, ognuno generalmente pari a alcuni MB. Tale suddivisione si presta bene per utilizzo con un filesystem parallelo per fornire prestazioni elevate. Generalmente usato per ambienti dove c'e' una lettura e scrittura di dati di medio-grandi dimensioni (diversi centinaia di blocchi).

Object storage

Salva dati in forma di oggetti dove sono inclusi anche i metadati. Generalmente e' usato per dati non strtturati e di varia tipologia. Particolarmente adatto in ambienti Cloud con workload molto eterogeneo. Associa quindi una elevata versalita' a buone prestazioni.

Filesystem: metodo che organizza i dati in strutture ad albero (directory e sotto directory)

NFS: Network Filesystem

NAS: Network Attached Storage

S3: Cloud Object Storage

https://www.cloudflare.com/learning/cloud/object-storage-vs-block-storage/

Concetti generali delle architetture storage scalabili: classificazione logica

What is a file system?
The file system is a fundamental of computing that allows data to be organised – usually in hierarchical directories – and retrieved. It is a logical system to help the operating system (OS) and user differentiate and organise information and also forms part of the physical addressing of data on storage media.
File systems specify conventions for file naming, such as filename length, which characters to use, case sensitivity, file type extension etc. A file system also keeps metadata about files, such as file size, creation date or location in the directory.
Most file systems organise files into a hierarchy, with file location described by a path within the directory structure. Directories are organised in an inverted hierarchical tree structure.
Physical media can be formatted to work with different file systems in partitions. Or, partitions can be created to help isolate files of different types from each other for performance or security reasons, such as OS files, user files and system files.
Partitions are divided into blocks devoted to, for example, file content, metadata and system data.
Access by users and applications is also controlled by the file system. That can be who has access to which files and directories as well as access control so that simultaneous writes cannot occur that might result in corruption or logical issues. Files can also be encrypted against external access.
A database management system (DBMS) is a little like a file system. But, whereas a file system provides interaction with the whole file and stores files as unstructured discrete items, a DBMS allows users to interact and change elements in a database near simultaneously. The DBMS manages the database as a consistent, single, highly controlled repository of data with robust security and access controls.
Block and file access storage offer two ways to interact with the file system

https://www.computerweekly.com/feature/Storage-technology-explained-File-block-and-object-storage?_gl=1*tvbo57*_ga*MTM4Mzk3NDgzMS4xNzQyNjU0MjQ2*_ga_TQKE4GS5P9*MTc0MjY3Mzc3Mi4yLjEuMTc0MjY3Mzc3NC4wLjAuMA..

https://www.cloudflare.com/learning/cloud/object-storage-vs-block-storage/

Concetti generali delle architetture storage scalabili: classificazione logica

What is file storage?
File storage, or file access storage, is storage in which entire files are accessed via the file system, usually via network-attached storage (NAS). Such products come with their own file system on board, from which storage is presented to applications and users in the drive letter format.
That contrasts with block storage, as we'll see below, and is a fundamental distinction in storage infrastructure.
File systems have numerous benefits. Among these is that most enterprise applications are written to interact with data via a file system, although that is being eroded by object storage (see below).
File storage accesses entire files, so is unstructured and suited to general file storage, as well as specialised workloads that require file access, such as in media and entertainment. In the form of scale-out NAS, it is a mainstay of large-scale repositories for analytics and high-performance computing (HPC) workloads.

What is block storage?
In block storage, storage-area network (SAN) hardware does not address entire files (although it can). Instead, block storage provides application access to the blocks of which files – in particular databases – are comprised.
This suits workloads where many users work on the same file simultaneously and from possibly the same application – email, enterprise applications such as enterprise resource planning (ERP), for example – but with locking at the sub-file level.
So, in the case of block storage, the file system through which applications talk resides higher in the stack, on host servers.
Block storage has the great benefit of high performance, and not having to deal with metadata and file system information

https://www.computerweekly.com/feature/Storage-technology-explained-File-block-and-object-storage?_gl=1*tvbo57*_ga*MTM4Mzk3NDgzMS4xNzQyNjU0MjQ2*_ga_TQKE4GS5P9*MTc0MjY3Mzc3Mi4yLjEuMTc0MjY3Mzc3NC4wLjAuMA..

https://www.cloudflare.com/learning/cloud/object-storage-vs-block-storage/

Concetti generali delle architetture storage scalabili: classificazione logica

What is object storage?

Object storage is the new kid on the block, relatively speaking.

Unlike file and block storage, it lacks a file system and is based on a “flat” structure with access to objects via their unique IDs. In this way, it’s similar to the domain name system (DNS) used to access web content.

So, object storage is not hierarchical, and lacks the directory system structure. That can be an advantage when datasets grow very large. Some NAS systems can become unwieldy when they get to billions of files.

Object storage also offers a richer set of metadata than traditional file systems, which makes it well-suited to data storage for analytics and artificial intelligence (AI).

Object storage accesses data in a way that looks more like file access, but it lacks the same kind of file locking. Often, for example, more than one user can access an object at the same time (think Google Docs). So, object storage is described as “eventually consistent”.

Most legacy applications are not written for object storage, but it is the storage access method of choice for the cloud era. That’s largely down to the fact that cloud object storage comprises the bulk of capacity offered by the hyperscaler cloud providers.

What is file, block and object storage in the cloud?

The cloud is the natural home of object storage, and it’s here that now de facto standards such as S3 emerged. Object storage is the bulk storage of the cloud era, and provides easy access to data that can happily exist as eventually consistent.

The big three hyperscaler cloud providers – Amazon Web Services (AWS), Microsoft Azure and Google Cloud Platform – also offer their own file and block storage services, as well as those from third-party storage suppliers.

Big-three cloud storage options include object storage such as S3 from AWS, Azure Blob and Google Cloud Storage.

File storage from the hyperscalers includes: Amazon’s Elastic File System (EFS), which is an NFS-based file system that operates on cloud and local storage; Azure Files, which uses SMB and allows concurrent file share mounting in the cloud or on-premise; and Google Cloud Filestore, which provides NAS for Google Compute Engine and Kubernetes Engines with storage offered at standard and premium levels.

Block storage from the big three comes as Amazon Elastic Block Store, which works with Amazon Elastic Compute Cloud; Azure Disk, which provides managed disks for Azure virtual machines; and Google Persistent Disk block storage, which runs up to 64TB, and offers standard persistent disks, persistent SSDs and local SSD.

All three hyperscalers also offer higher-performing file storage based on NetApp storage. Pure Storage Cloud Block Store is available on AWS

Concetti generali delle architetture storage scalabili: classificazione logica

What's the difference between file locking and object locking?

A fundamental function of file systems is their locking mechanisms. These make sure different users and applications that work on the same file simultaneously cannot cause conflicts that result in inaccuracies and inconsistencies in the data.

Locking is strong and well-developed in file systems. However, object storage is not built around a file system, so it lacks the same kind of methods that enable locking.

File (NAS) and block (SAN) storage both rest on the file system. NAS storage accesses files directly, while block storage accesses blocks in the file system to update parts of a database, for example, which itself comprises a "file".

Windows systems can set file locking by application and user for whole files to restrict access, shares, reads, writes and deletes, or byte-range locks for regions of files.

Unix-like file systems, including Linux, vary between the distributions, but you can modify open files in Linux, for example. Differences are to do with how Windows and Unix-like systems record file information, but they can all restrict access and changes to files.

Meanwhile, object storage lacks built-in locking. It's not that it doesn't exist in object storage, but it's not built into object storage in the same way as it is with file systems. Multiple users can work on the same object at once, with changes reconciled on an "eventually consistent" basis. Some forms of locking are implemented in object storage and the cloud. These include file access protocol gateways that sit in front of object stores.

What's the difference between NFS, SMB and CIFS?

NFS, SMB and CIFS are all file storage protocols that give access to files on servers and storage servers (such as NAS storage) as if they were local files.

They are distinct from the file system, being protocols that operate at the application layer to facilitate communication between applications and storage, via the file system. They are application layer protocols, of the same order as HTTP, FTP, POP and SMTP, for example.

NFS, SMB and CIFS are used with NAS file access storage, not SAN block access storage.

NFS is mostly used with Linux and Unix operating systems, and was originally developed by Sun Microsystems in 1984. It reached version 4.2, with parallel file access functionality (pNFS, used in scale-out NAS), in 2016.

Although developed by a Unix supplier and often used for Unix and Linux, NFS can also be used in Windows environments.

SMB is primarily used in Windows environments, and is the basis for Microsoft's distributed file system. IBM first developed SMB in 1983 to provide shared network access to files and printers. Microsoft picked it up later and built it into Windows NT 3.1. It has retained it in its operating systems since then.

CIFS is an implementation of SMB, first introduced in 1996. It's mostly used with NetBIOS-based transports and was focused on small LAN file, print and application access to storage. It's less scalable than NFS, and considered chatty, buggy and less secure than SMB.

Concetti generali delle architetture storage scalabili: classificazione logica

What are the pros and cons of object storage and block storage?

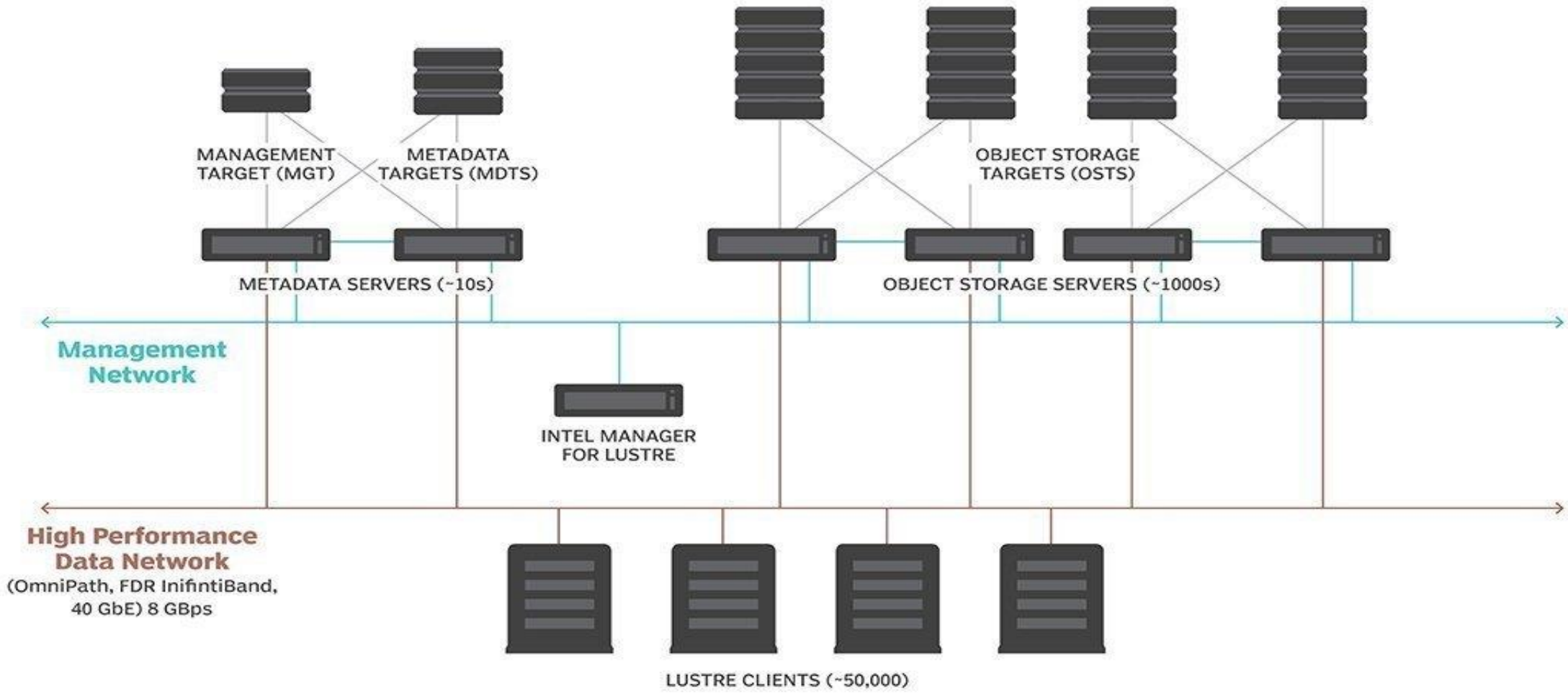
Capability	Block storage	Object storage
Storage capacity	Limited	Nearly unlimited
Storage method	Data stored in blocks of fixed size, reassembled on demand	Unstructured data in non-hierarchical data lake
Metadata	Limited	Unlimited and customizable
Data retrieval method	Data lookup table	Customizable
Performance	Fast, especially for small files	Depends, but works well with large files
Cost	Depends on vendor, usually more expensive	Depends on vendor, usually less expensive (aside from egress fees)

https://www.cloudflare.com/learning/cloud/object-storage-vs-block-storage/

Esempio di filesystem paralleli e di architetture storage

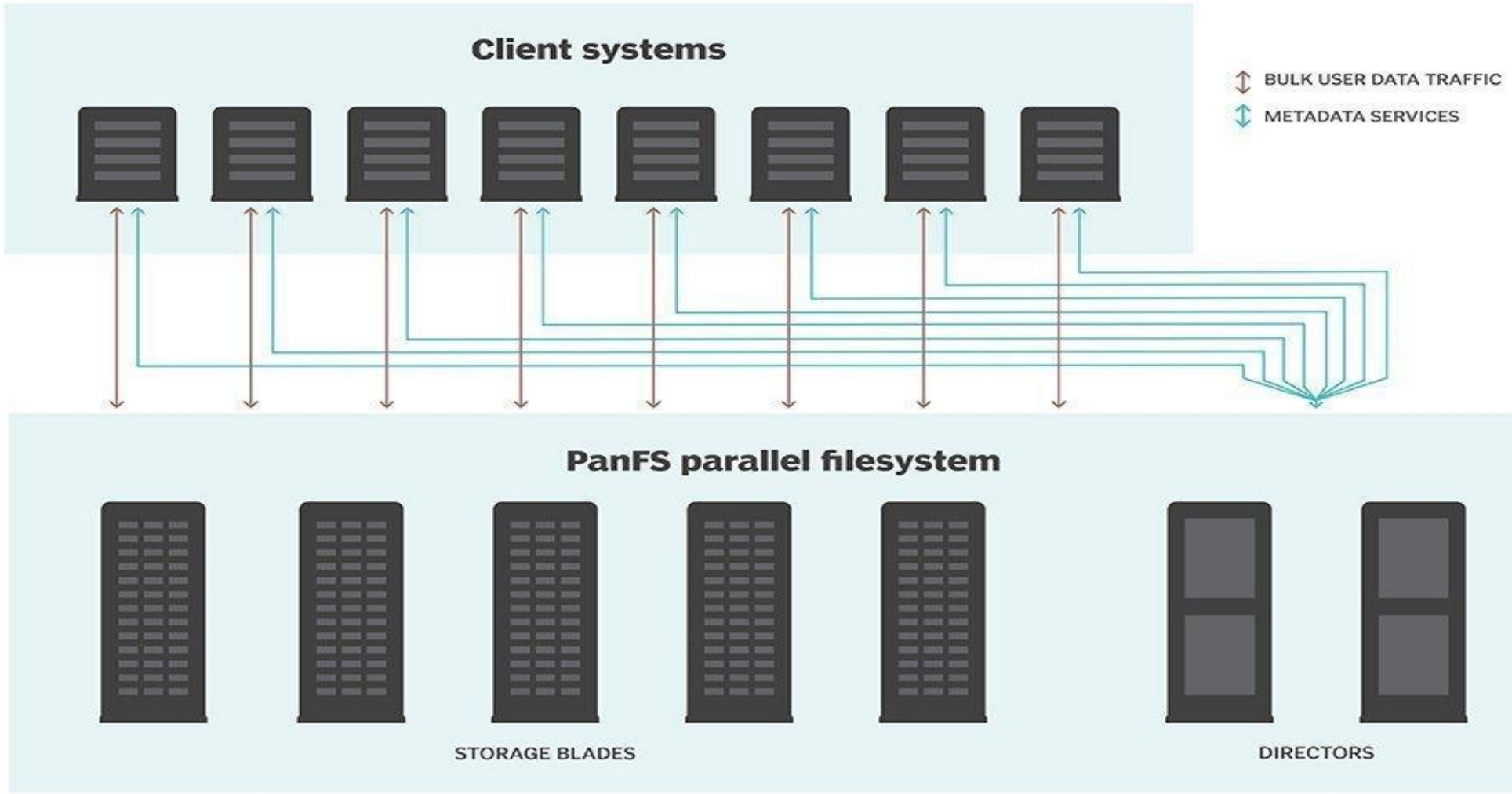
https://cloudian.com/blog/scale-up-vs-scale-out-storage/

Lustre architecture



SOURCE: INTEL; ICONS: MACROVECTOR/GETTY IMAGES

Panasas storage architecture



SOURCE: PANASAS; ICONS: MACROVECTOR/GETTY IMAGES

IBM Storage Scale – Overview

What is Storage Scale?

- Highly scalable distributed parallel POSIX file system
- Runs on any Open System OS (AIX, Windows, Linux) environment
- Utilizes any block storage device (internal media or SAN storage)

Under the file system

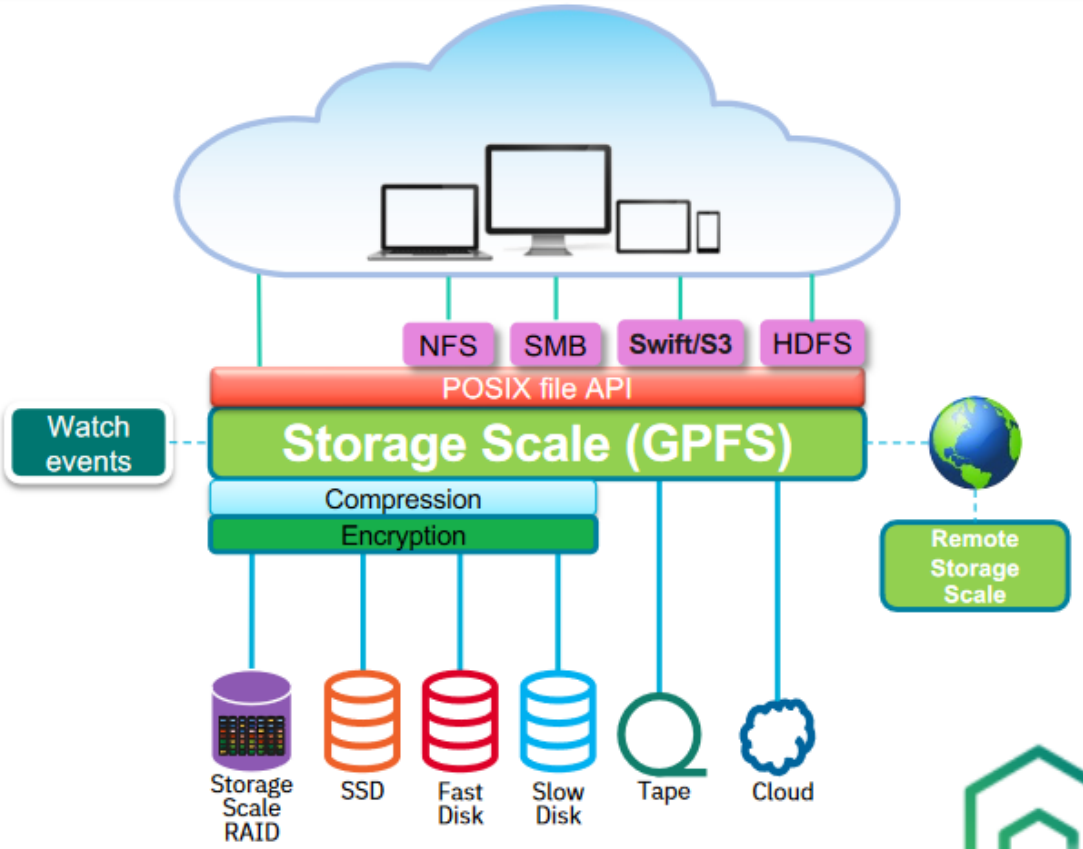
- Use any block storage usable by the OS.
- Information Lifecycle Management features enable tiering
- Built-in encryption and compression
- Tape and cloud-based storage usable as tiers.
- Storage Scale RAID gives unrivaled data integrity.

Above the file system – since not everything will run Scale directly…

- Export the data through multiple protocols (SMB,NFS, Object, HDFS)
- Container Storage Interface (CSI) extends storage access to containers.

Beside the file system

- Peer with remote Storage Scale file systems for efficient remote access.
- Replication capabilities for robust Disaster-Recovery capabilities.
- Watch events in the file system with Watch Folders and File Audit Logging.



In a single name space, vastly simplify administration and reduce costs by eliminating storage islands.

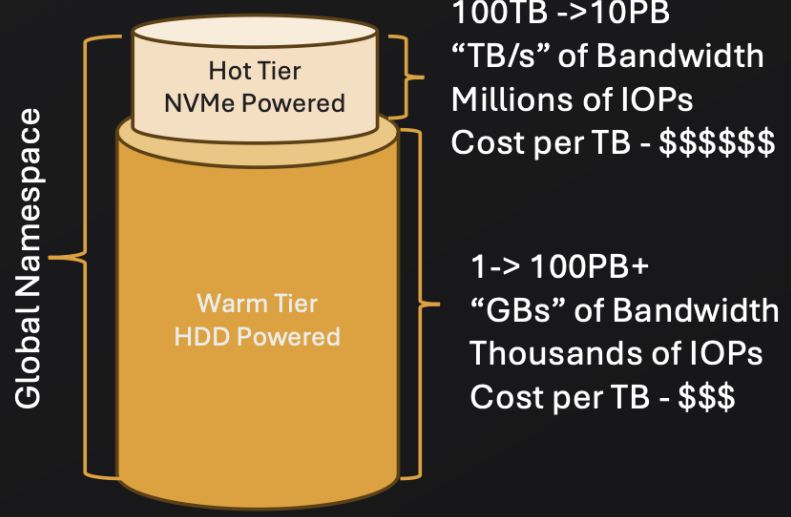
Un esempio di soluzione storage ibrida a multilivelli

VDURA Hybrid Advantage

Other Vendors (~~Weka~~)

- Only support the software on the HOT Tier
- Does not support the Warm Tier or any HDD HW
- May recommend low performance warm tiers and super small ratios of flash
- Complex configurations and tools for Tiering

Hybrid Storage Design , Most Target 5-20% Flash



VDURA Advantage

- Single Vendor support for End 2 End Storage Solution (HW+SW).
- Provide Healthy Ratios of Flash/HDDs
- Best in Class High Bandwidth Parallel IO HDD tier .
- 100% transparent data Tiering managed by VDURA
- Best in class Durability and Avaibility

https://www.vdura.com/

Benchmark generalmente utilizzati per architetture storage a alte prestazioni

IOR: (Interleaved or Random) e' comunemente usato per misurare le prestazioni file systems paralleli.
https://wiki.lustre.org/IOR

IOZONE: simile a IOR ma ormai meno usato

IOPS: Input/Otput operations per second molto usato per definire le prestazioni
in lettura e scrittura anche di dati di piccole dimensioni. Molto usato per sistemi flash.

Measurement	Description
Total IOPS	Total number of I/O operations per second (when performing a mix of read and write tests)
Random Read IOPS	Average number of random read I/O operations per second
Random Write IOPS	Average number of random write I/O operations per second
Sequential Read IOPS	Average number of sequential read I/O operations per second
Sequential Write IOPS	Average number of sequential write I/O operations per second

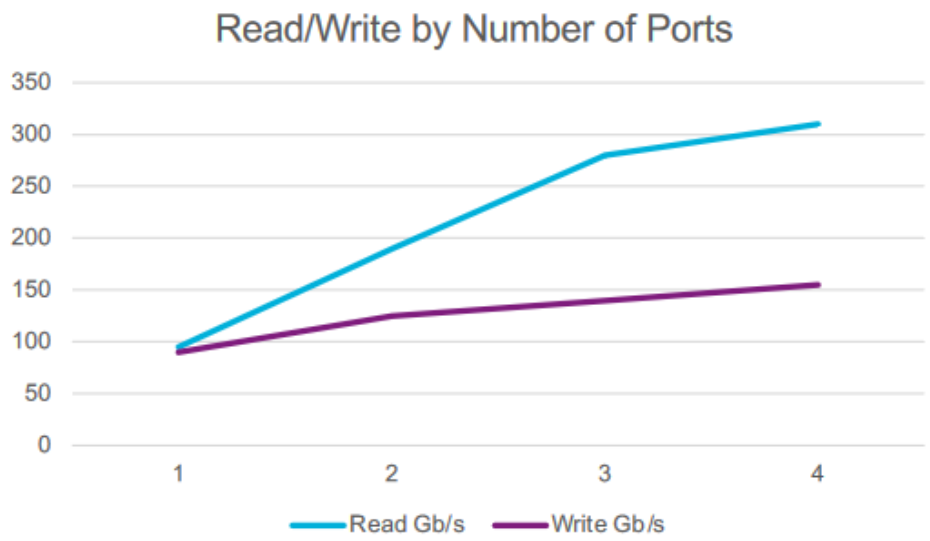
IO500: un benchmark composito che misura IO per dati non strutturati di varie dimensioni
in sistemi paralleli https://io500.org/

Comparative Performance Numbers – Varying Bandwidth

Using:

- Scaling out the Infiniband Adapters
- Sequential Bandwidth
- 8+2p RAID
- 16MB filesystem block size

SSS 6000 IB adapters (2x 200Gb ports each)	Read GB/s	Write GB/s
1	95	90
2	190	125
3	280	140
4	310	155

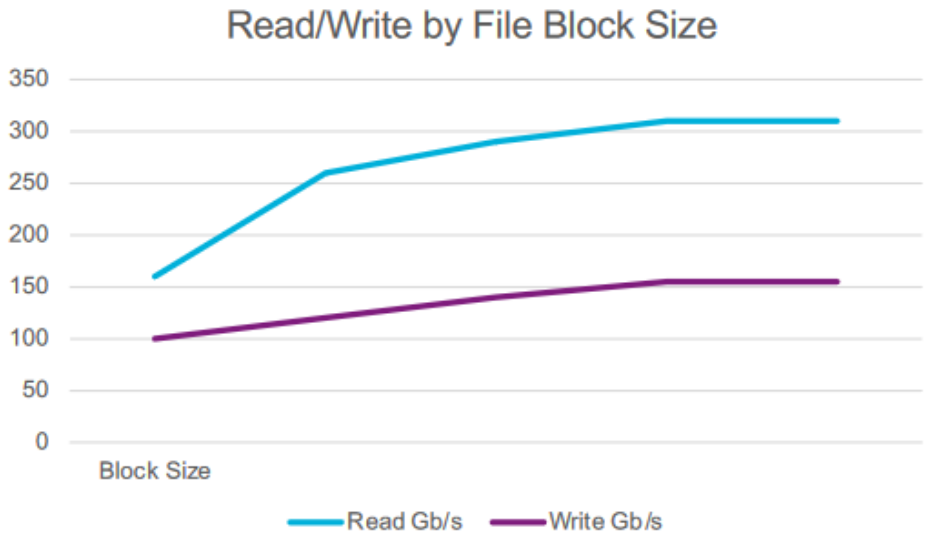


Comparative Performance Numbers – Various Block Sizes

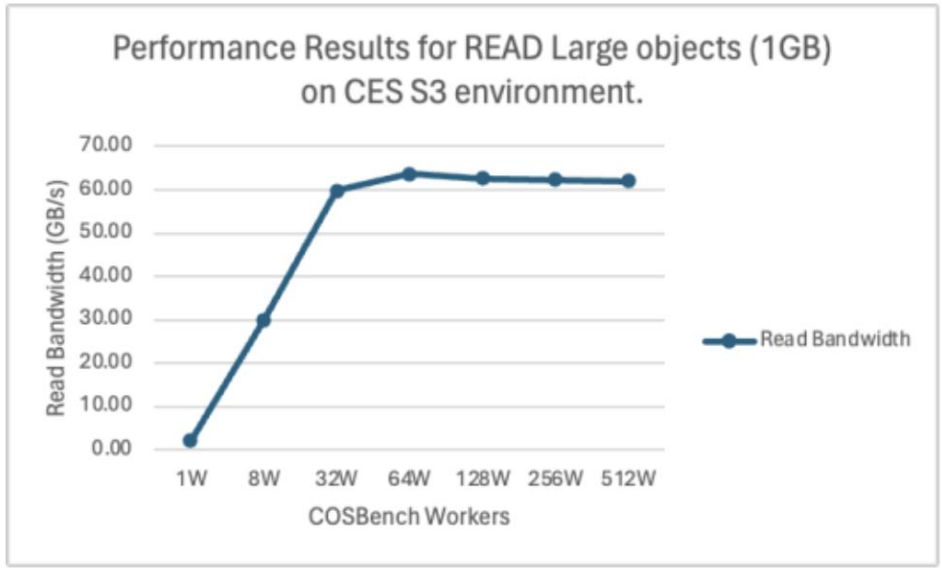
Using:

- Varying filesystem block sizes
- Sequential Bandwidth
- A “maxed out” Config
- 8+2p RAID
- 16MB filesystem block size

SSS 6000 Filesystem Block Size(MB)	Read GB/s	Write GB/s
1	160	100
2	260	120
4	290	140
8	310	155
16	310	155

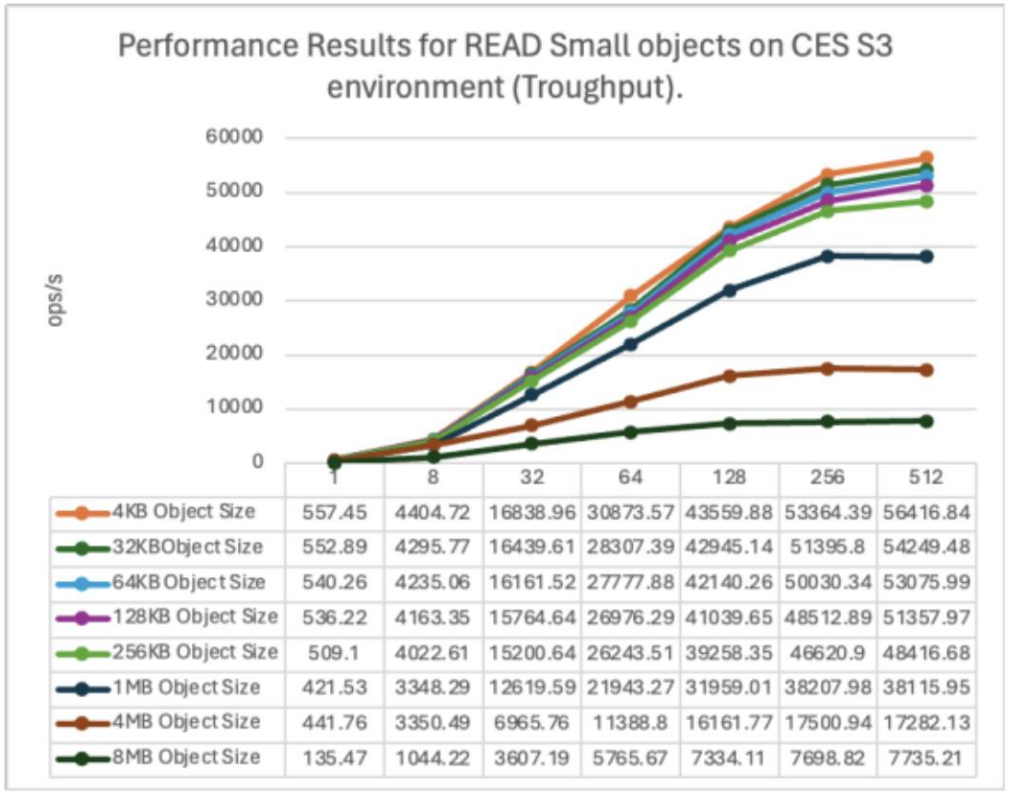


Access Services – Object Performance



Op-Type	Obj Size	Workers	Op-Count	Byte-Count	Avg-ResTime	Avg-ProcTime	Throughput	Bandwidth	Succ-Ratio
READ	1GB	1	611 ops	625.66 GB	490.86 ms	4.83 ms	2.04 op/s	2.09 GB/S	100%
		8	8.78 kops	8.99 TB	273.26 ms	5.13 ms	29.27 op/s	29.98 GB/S	100%
		32	17.49 kops	17.91 TB	548.53 ms	7.67 ms	58.33 op/s	59.73 GB/S	100%
		64	18.61 kops	19.06 TB	1029.76 ms	15.79 ms	62.15 op/s	63.64 GB/S	100%
		128	18.28 kops	18.72 TB	2093.59 ms	28.6 ms	61.13 op/s	62.6 GB/S	100%
		256	18.12 kops	18.55 TB	4210.39 ms	60.39 ms	60.79 op/s	62.25 GB/S	100%
		512	17.91 kops	18.34 TB	8453.23 ms	106.39 ms	60.55 op/s	62.01 GB/S	100%

Table 2. Performance Results for READ Large objects (1GB) on CES S3 envronment.

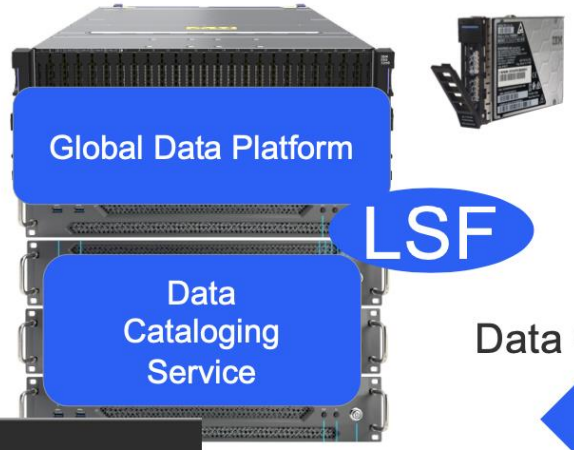


https://community.ibm.com/community/user/storage/blogs/rogelio-rivera-gutierrez/2024/04/25/ibm-storage-scale-performance-ces-s3-tech-preview

Data Orchestration providing a future scalable architecture with Tape - IBM S3 Deep Archive

https://www.ibm.com/downloads/cas/84KXPK05

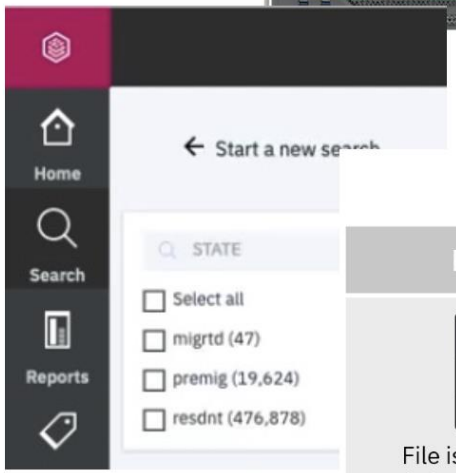
IBM Storage & SDI

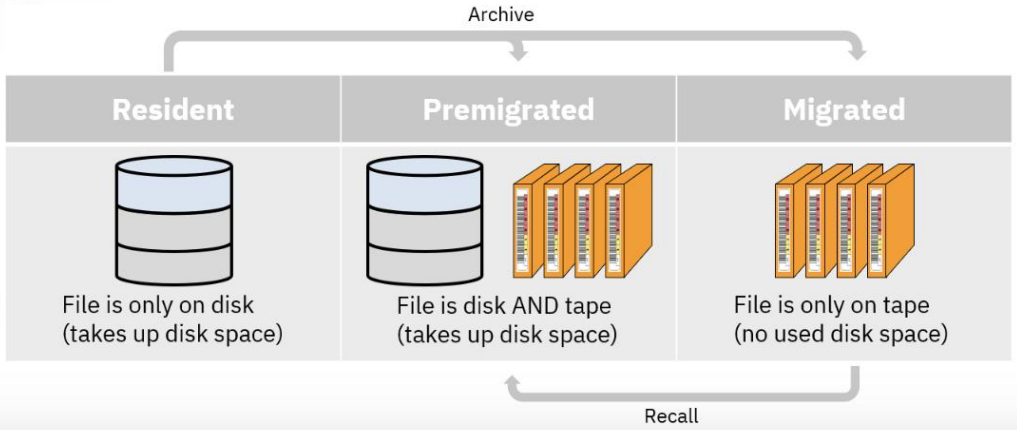


Architecture to ingest and recall data to tape libraries

Data Orchestration Services

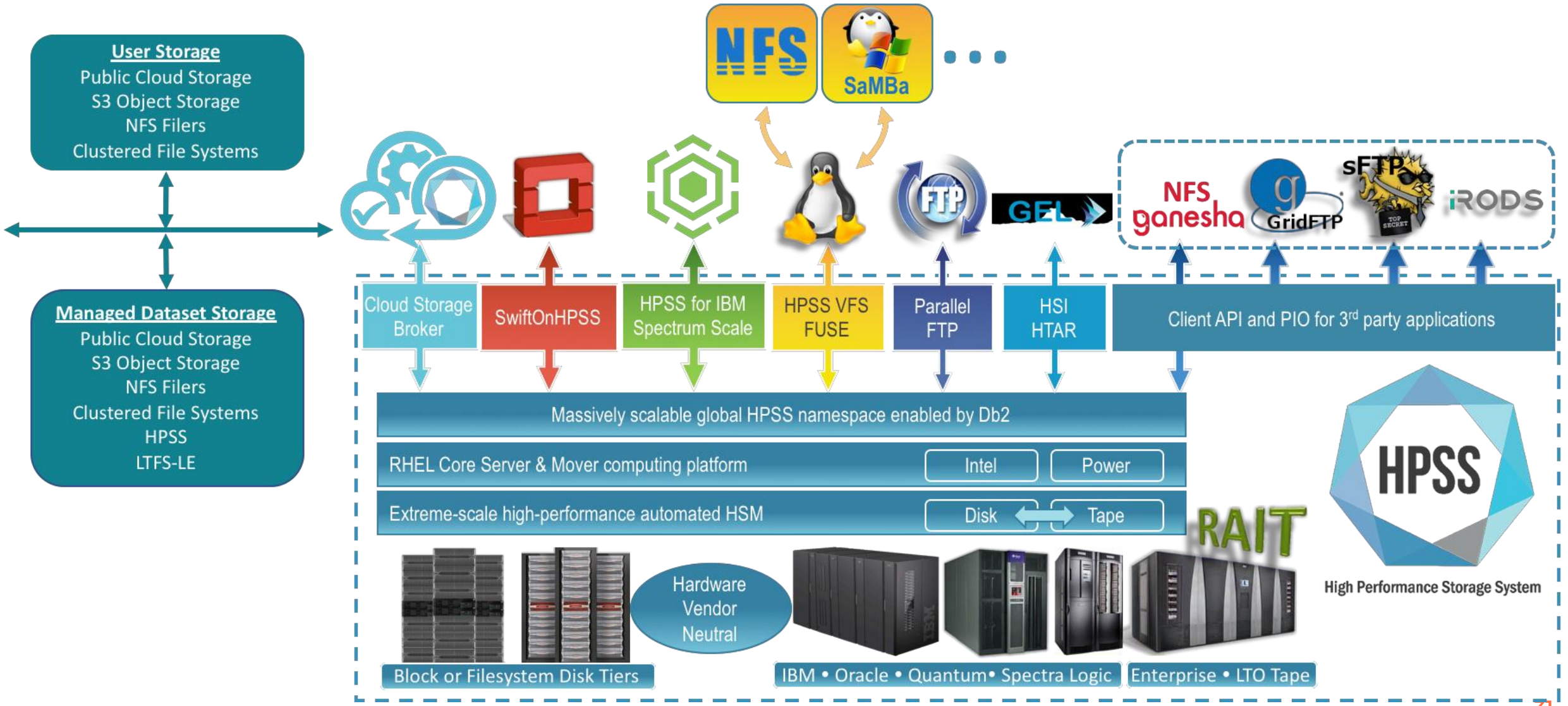
Orchestration





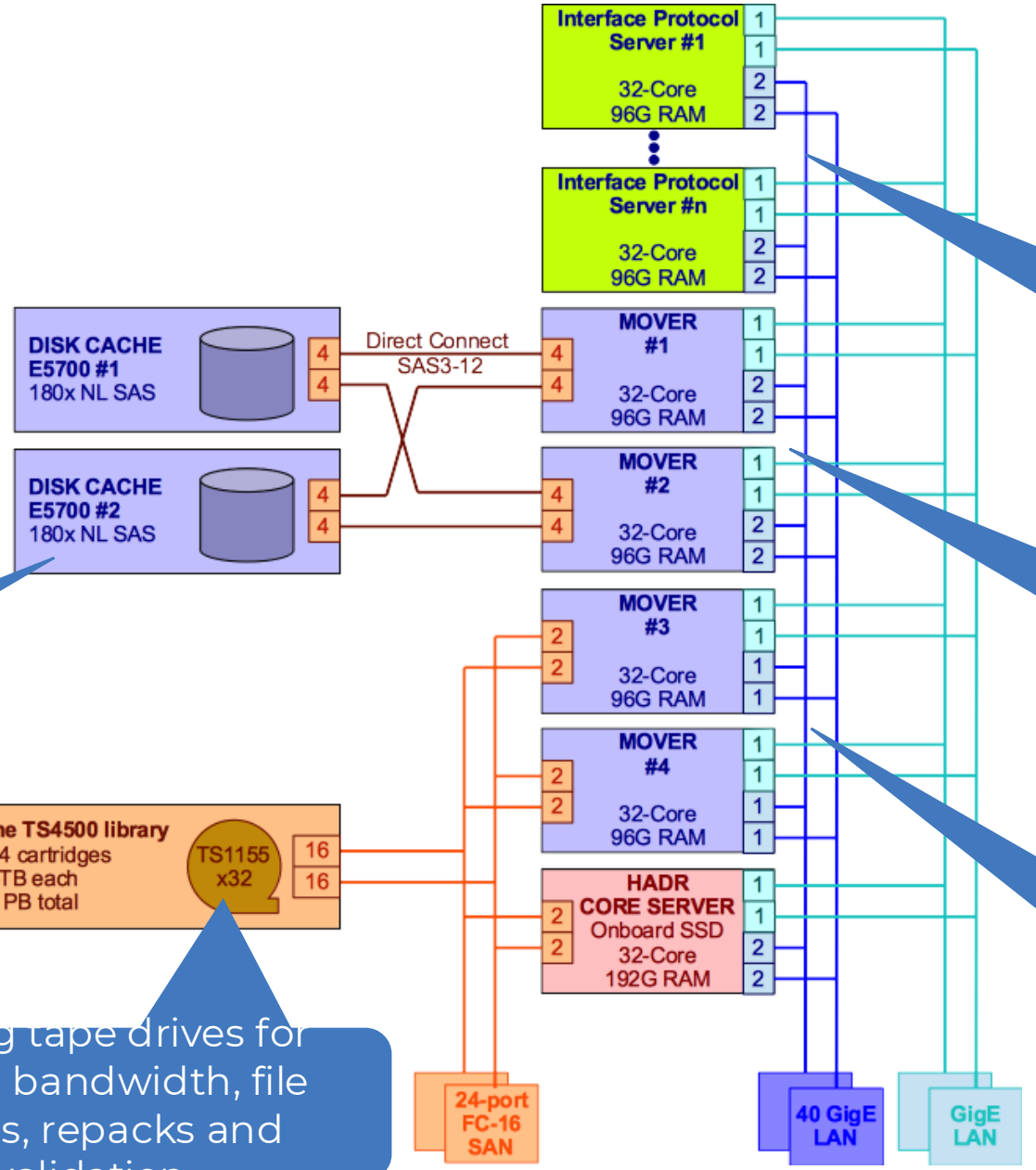
TS4500 or Diamondback Racks

The HPSS Big Picture



Scaling HPSS

eXtREMELy SCaLaBLE



Adding storage units for more disk cache capacity and bandwidth

Adding libraries for higher mount rates and higher capacity

Adding tape drives for higher bandwidth, file recalls, repacks and validation

Adding interface protocol servers for more client connections

Adding movers for increased tape drives and disk cache storage units

Dedicated Tape Movers with 32-cores support 16 tape drives

Piano del corso – 6th lesson

Problematiche di efficientamento energetico per sistemi HPC a grande scala (architetture pre e exascale)

Il concetto di PUE e di efficienza energetica a parita' di potenza computazionale

Come le varie architetture si caratterizzano in termini di "Potenza di Calcolo"/Watt

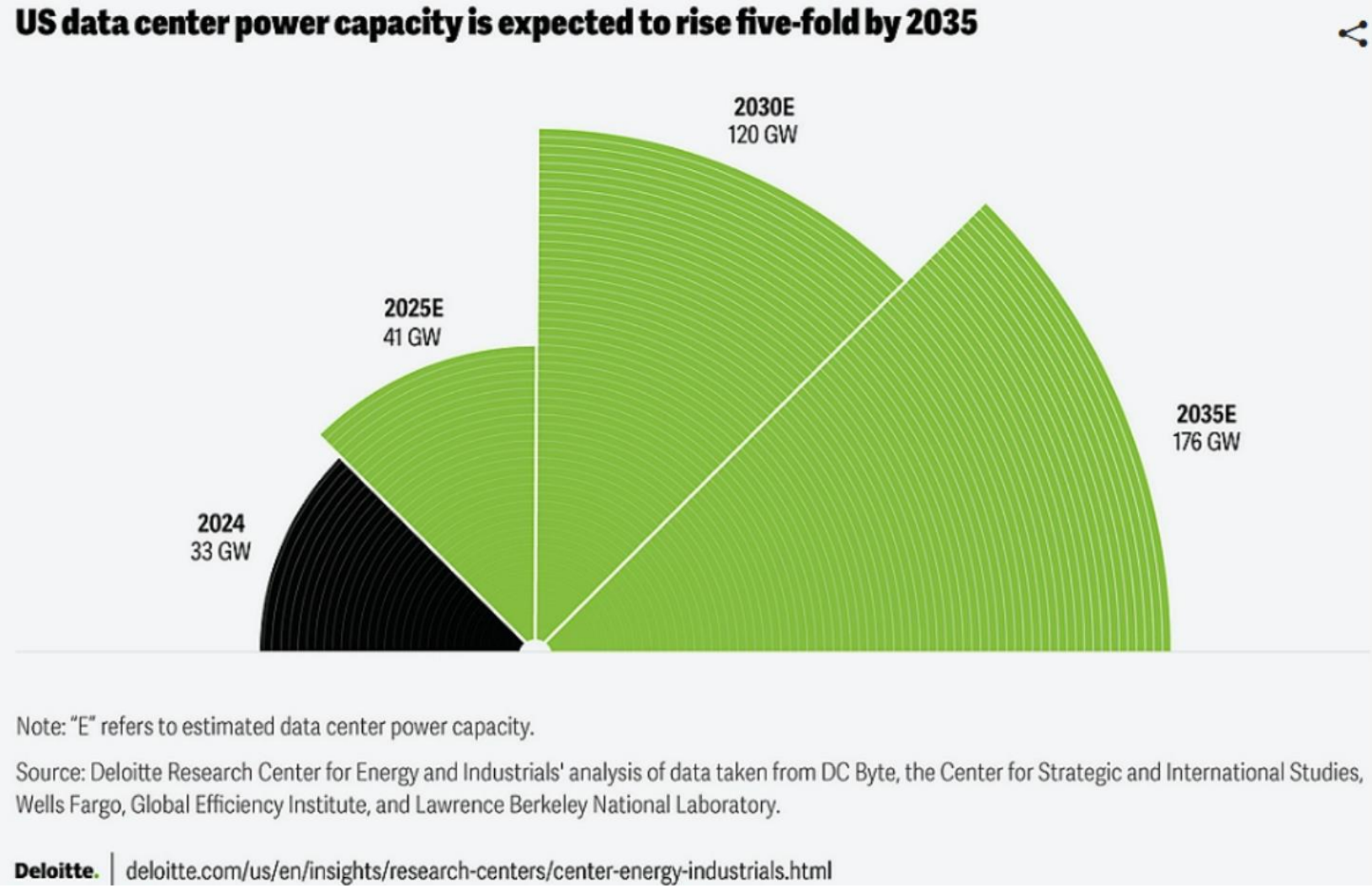
Utilizzo di tecniche di gestione del carico di lavoro per ottimizzare l'efficienza energetica

Soluzioni di raffreddamento a aria, a acqua diretta

Concetti generali sul disegno e la realizzazione di Data Center efficienti

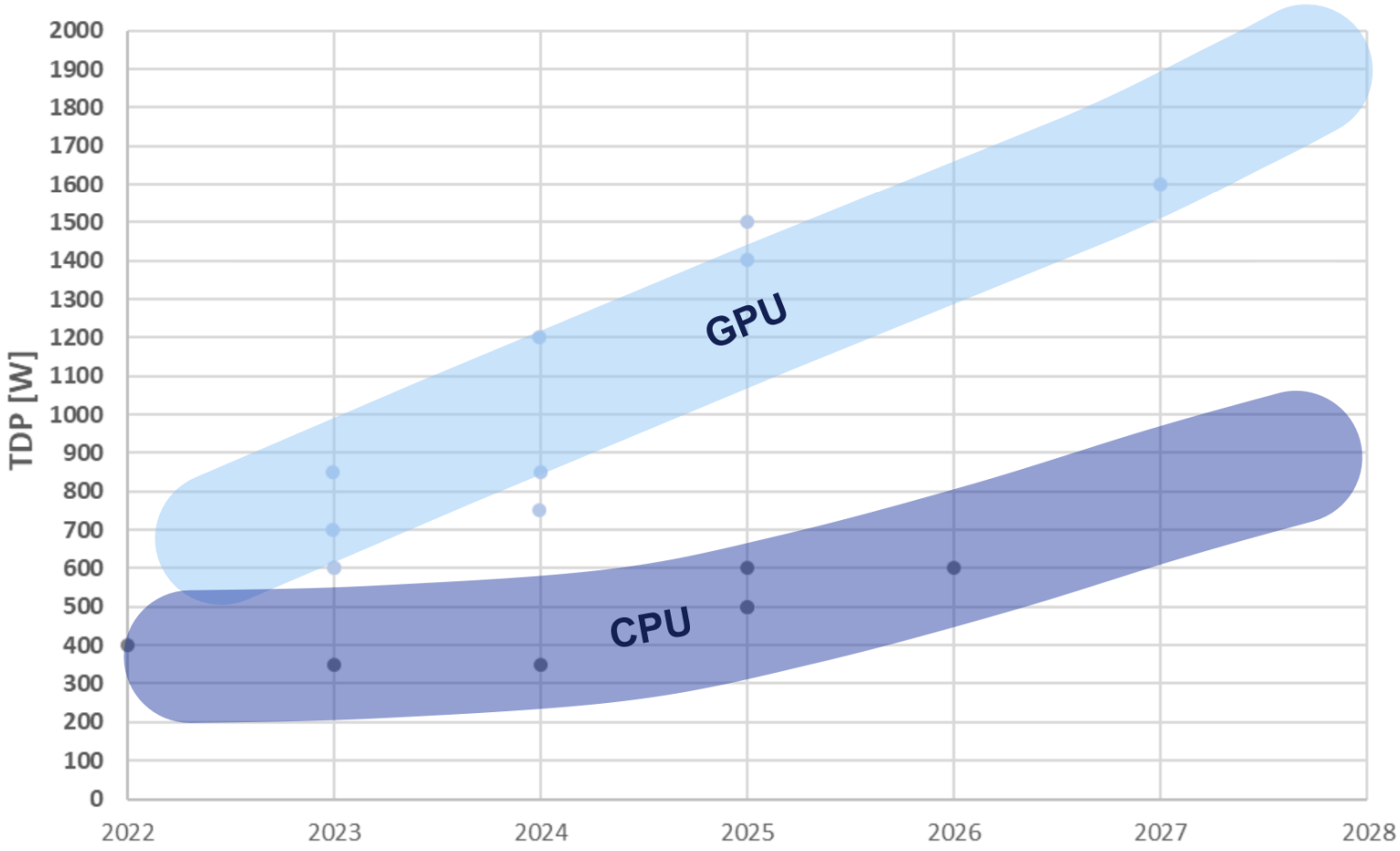
Data Center power capacity – future trends

Deloitte Reports on Nuclear Power and the AI Data Center Energy Gap

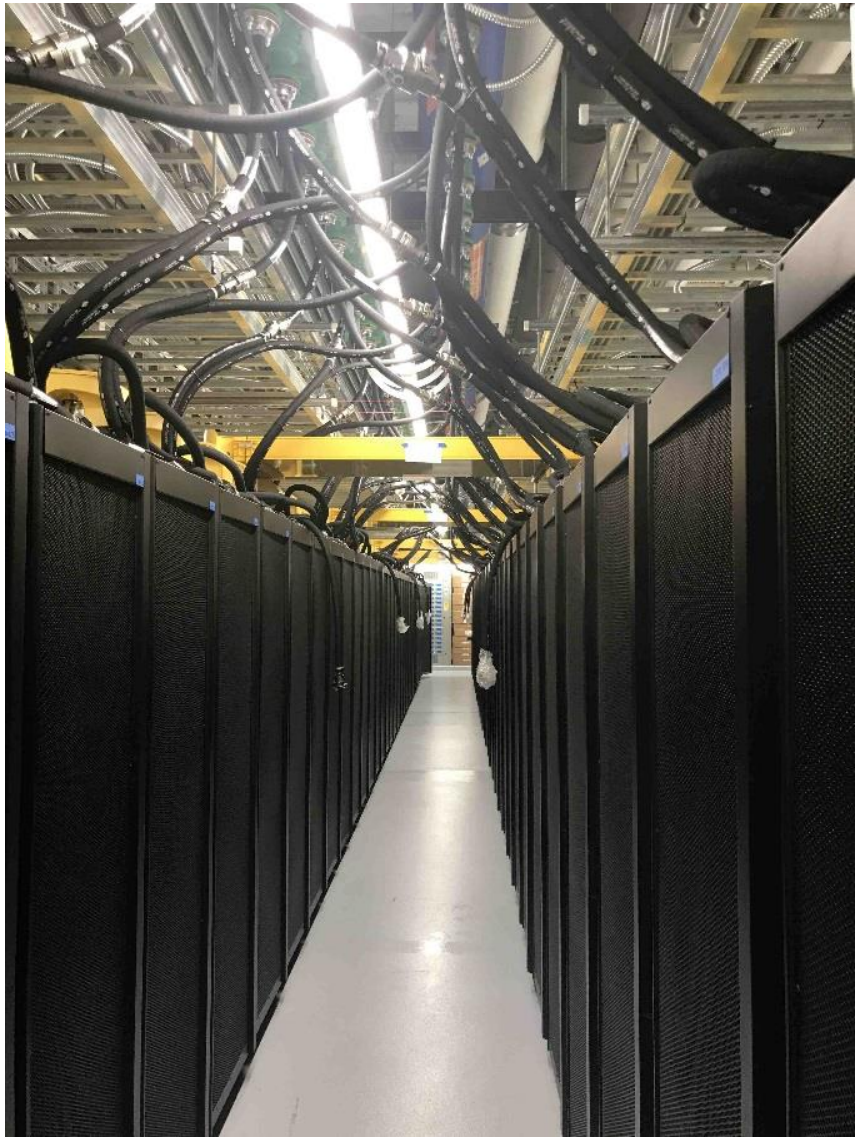


https://insideainews.com/2025/04/18/deloitte-reports-on-nuclear-power-and-the-ai-data-center-energy-gap/

CPU vs GPU TDP – future trends



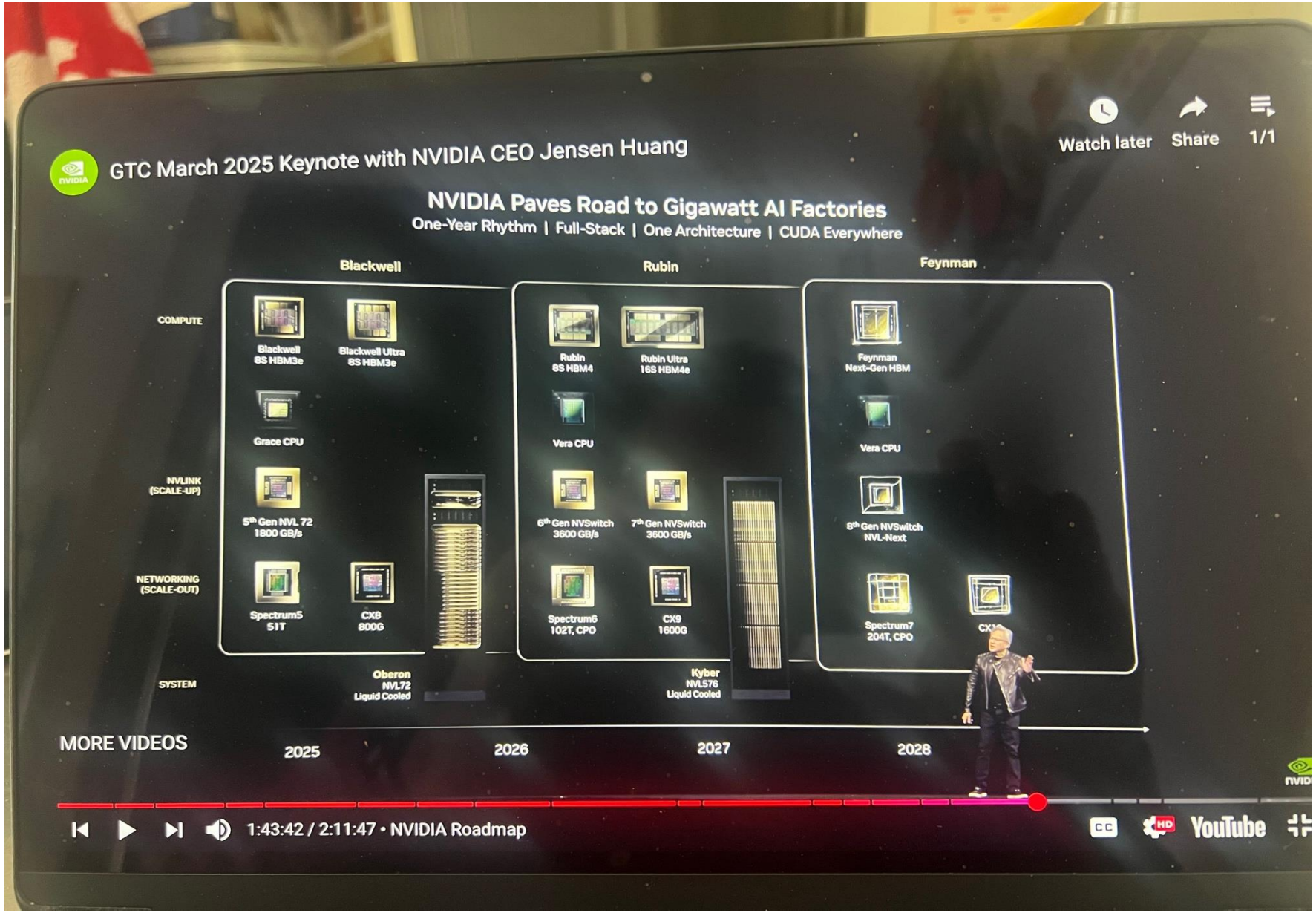
CORAL: ORNL‘s Summit System



Overhead H_2O distribution

Supercomputer Leonardo al Tecnopolo di Bologna

NVIDIA road-map



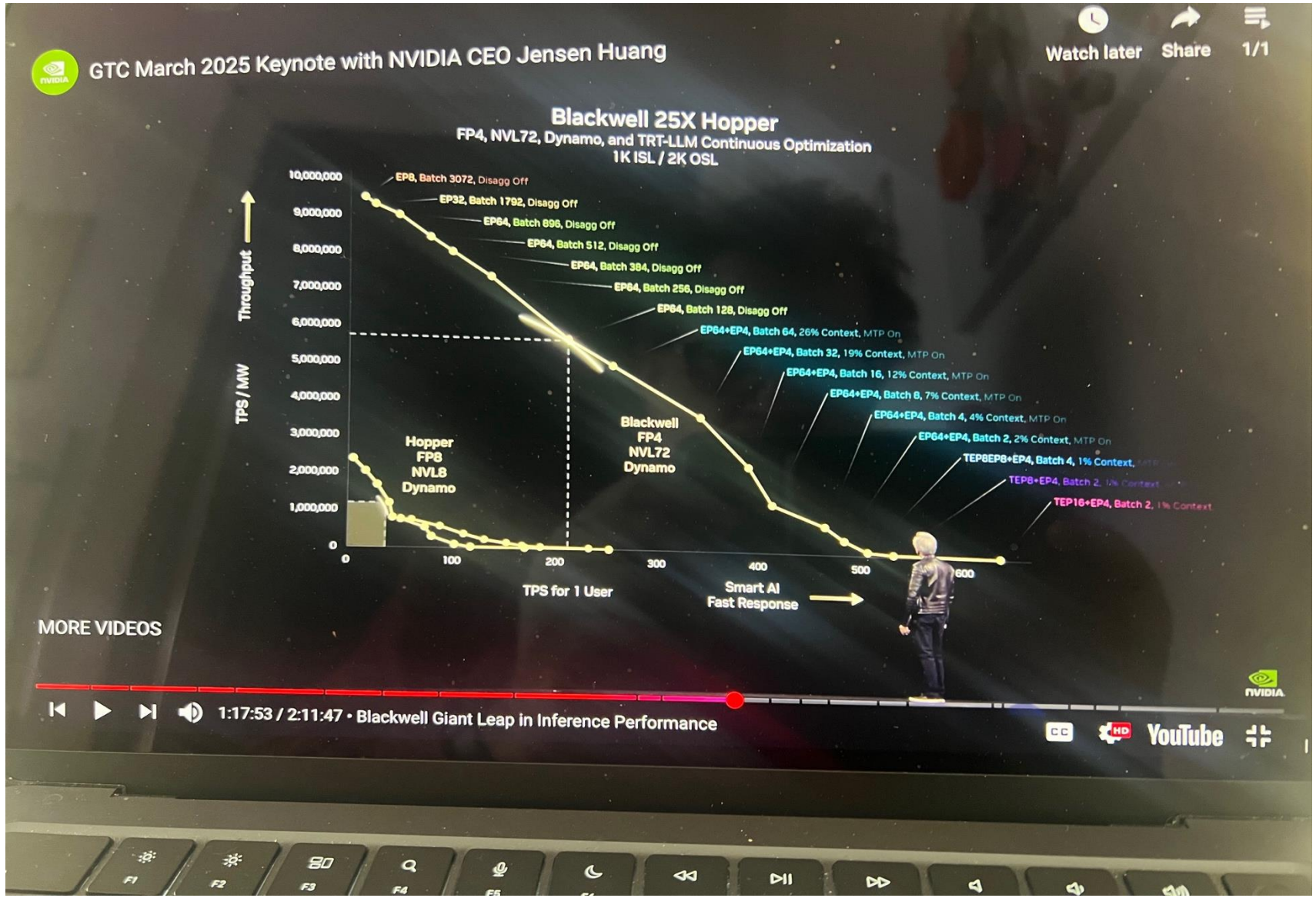
GPU TDP specification

Blackwell: 1200W max power

Rubin: 1500W max power est

https://www.nvidia.com/gtc/keynote/ https://www.theregister.com/2024/03/18/nvidia_turns_up_the_ai/

Token per second vs MW in AI factory



In economics, the **Jevons paradox**; sometimes **Jevons effect**) occurs when technological advancements make a resource more efficient to use (thereby reducing the amount needed for a single application); however, as the cost of using the resource drops, if the price is highly elastic, this results in overall demand increases causing total resource consumption to rise.[1][2][3][4] Governments have typically expected efficiency gains to lower resource consumption, rather than anticipating possible increases due to the Jevons paradox.[5]

https://www.nvidia.com/gtc/keynote/

https://en.wikipedia.org/wiki/Jevons_paradox

Processori e loro limiti alla scalabilita’ prestazionale

Theorethical Peak Perf = clock * fops * cores

Generalmente fops = 16 ÷ 32 64bits (Intel e AMD (recenti annunci))
iops = 32 ÷ 64 32bits

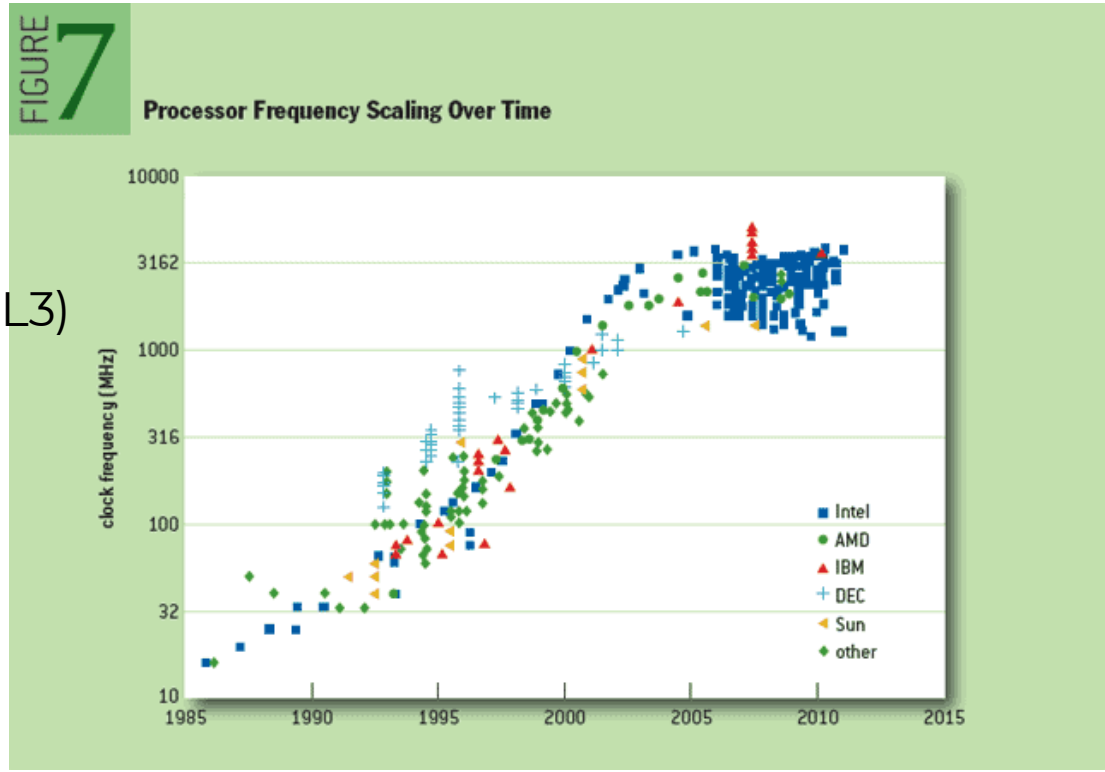
La frequenza del processore non puo’ crescere oltre un limite per i consumi e la dissipazione di calore

Attuali clock ≤ 4GHz → TDP ≤ 450°C

Aumento del numero di cores per CPU per aumentare le prestazioni

Limite al numero di cores dovuti alla memory BW (L2 e L3)

Il clock non e’ univoco: AVX per alcune architetture potrebbe avere un clock circa 70% del valore nominale



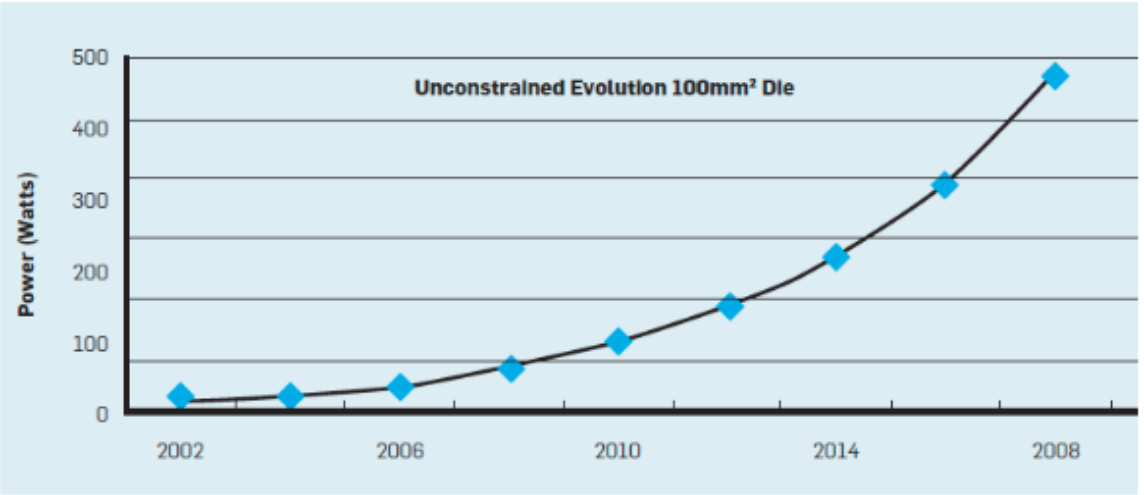
Review: Moore's Law

- **Empirical observation**
  - —transistor count doubles approximately every 24 months
    - features shrink, semiconductor dies grow
- **Impact: performance has increased 1000x over 20 years**
  - —microarchitecture advances from additional transistors
  - —faster transistor switching time supports higher clock rates

John Mellor-Crummey
Department of Computer Science
Rice University

Unconstrained Evolution vs. Power

- If
  - —add more cores as transistors and integration capacity increases
  - —operate at highest frequency transistors and designs can achieve
- Then, power consumption would be prohibitive



- Implications
  - —chip architects must limit number of cores and frequency to keep power reasonable
    - severely limits performance improvements achievable!

Figure credit: Shekhar Borkar, Andrew A. Chien, The Future of Microprocessors. Communications of the ACM, Vol. 54 No. 5, Pages 67-77 10.1145/1941487.1941507.

Limiti nella densita‘ dei mictoprocessori rispetto ai consumi elettrici e alla dissipazione del calore

Starting in 2021, having reached the limit of horizontal integration at the 40 nm / 16 nm (gate pitch / metal pitch) level, CMOS logic manufacturers will begin to stack several layers of transistors on top of each other with up to 6 layers by 2037 with the 3DVLSI technology.

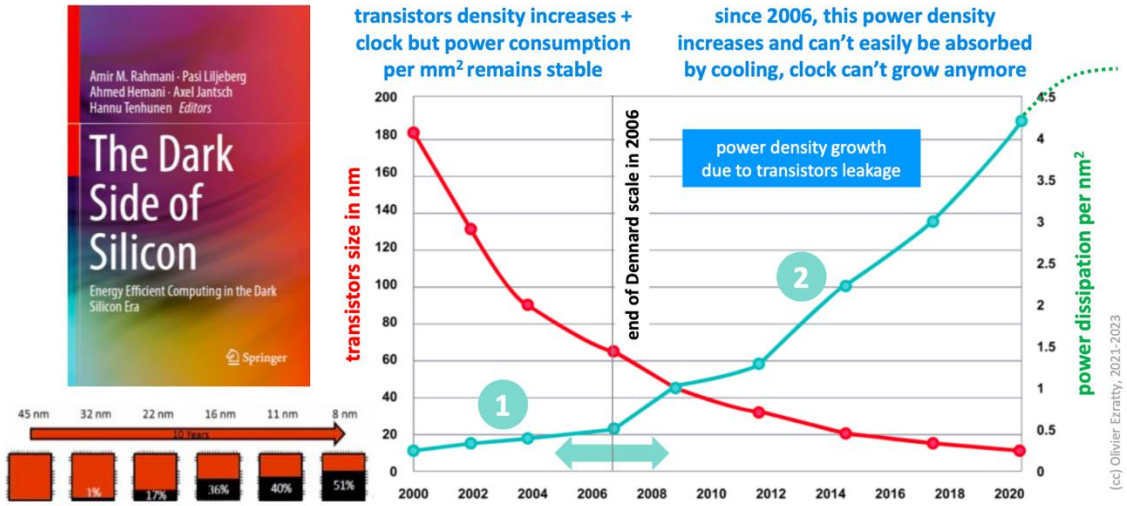


Figure 3: Dennard's scale and the dark side of silicon. Various sources like Hennessy[18].

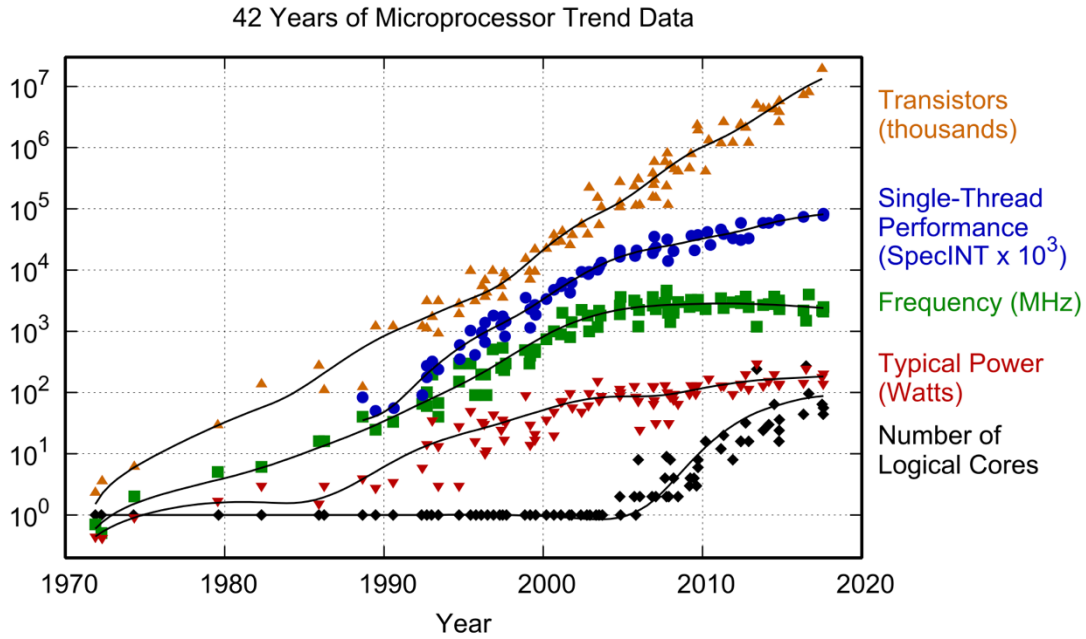


Figure 4: how microprocessor figures of merits progress slowed down with single thread performance, clock speed, and number of logical cores, in relation with total power consumption. Still, Moore's law related to the number of transistors per chipset is always valid. Source: Karl Rupp[32].

Is there a "Moore's law" for quantum computing?

Olivier Ezratty 1

Server Power Trends – **ASHRAE* 2015-2020**

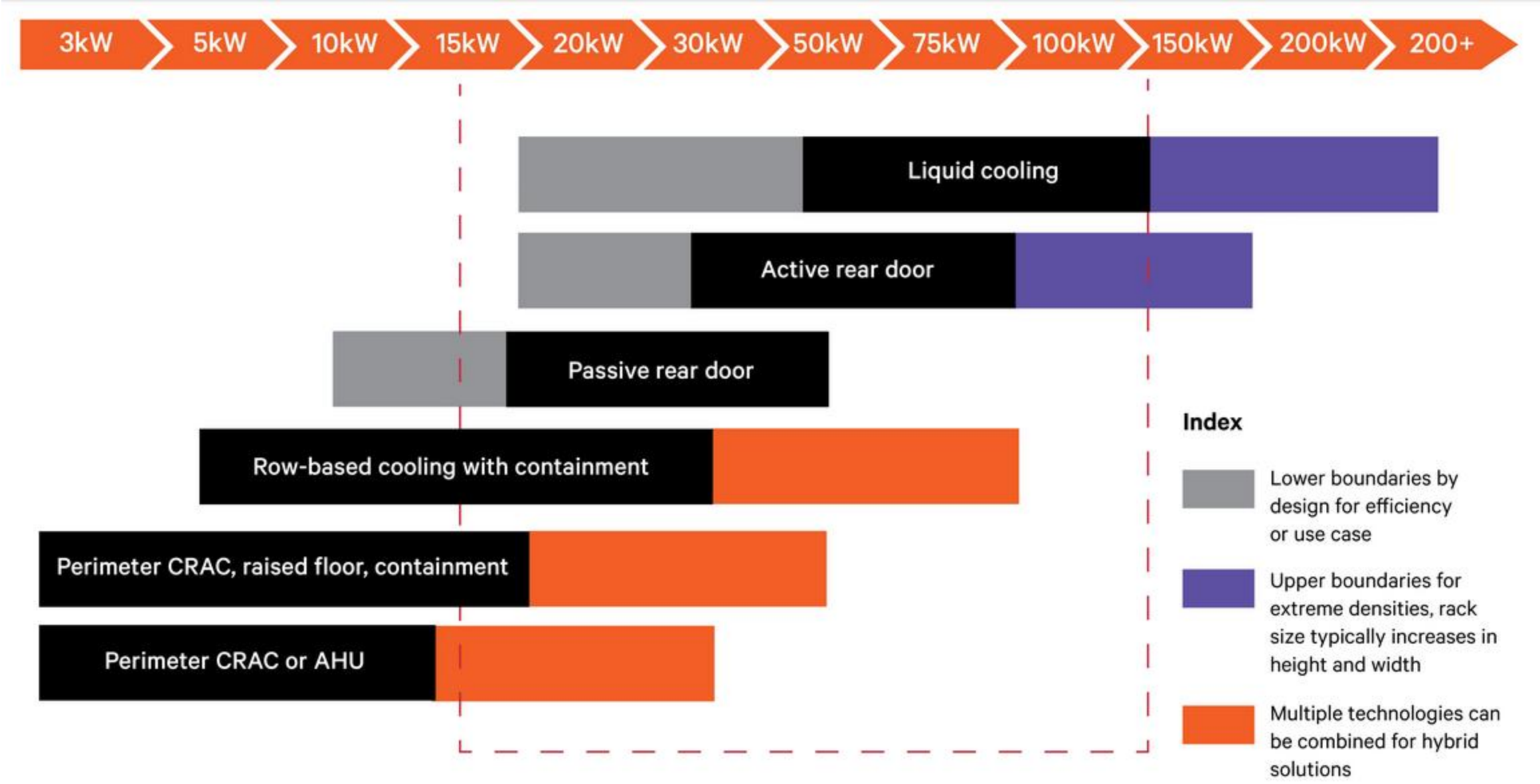
Market Requirements force IT manufacturers to maximize performance/volume creating high heat load/rack

Height	No. of Sockets	Heat Load / Chassis (watts)			Heat Load / 42U Rack			Increase 2010 to 2020
		2010	2015	2020	2010	2015	2020	
1U	1s	255	290	330	10,710	12,180	13,860	29%
	2s	600	735	870	25,200	30,870	36,540	45%
	4s	1,000	1,100	1,200	42,000	46,200	50,400	20%
2U	2s	750	1,100	1,250	15,750	23,100	26,250	67%
	4s	1,400	1,800	2,000	29,400	37,800	42,000	43%
4U	2s	2,300	3,100	3,300	23,000	31,000	33,000	43%
7U (Blade)	2s	5,500	6,500	7,500	33,000	39,000	45,000	36%
9U (Blade)	2s	6,500	8,000	9,500	26,000	32,000	38,000	6%
10U (Blade)	2s	8,000	9,000	10,500	32,000	36,000	42,000	31%

These rack heat loads will result in increased focus on improving data center ventilation solutions and localized liquid cooling solutions

*ASHRAE = American Society of Heating, Refrigerating, and Air-Conditioning Engineers. The group provides operating environment standards for datacenter operations.

Differenti sistemi di raffreddamento in funzione del carico termico sul rack



https://www.vertiv.com/it-emea/solutions/learn-about/liquid-cooling-options-for-data-centers/

Soluzioni tecnologiche per il raffreddamento dei server

- Capacita' estrattiva del calore prodotto: $\backsim$ C x Portata in regime laminare x ΔT

$$C_{Aria} = 0{,}24\ Cal/g \times C \quad - \quad C_{H2O} = 1\ Cal/g \times C \quad - C_{H2O} / C_{aria} \backsim 4$$

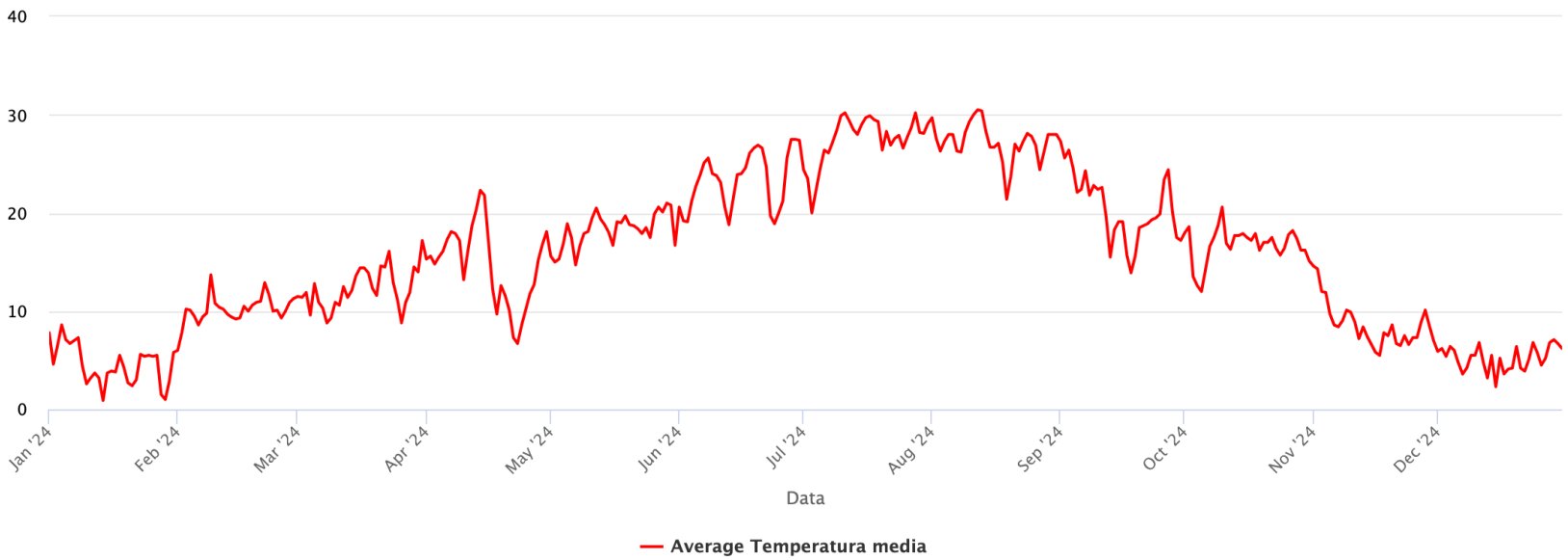
- Raffreddamento a aria – generalmente aria d'ingresso tra 25° e 32°C
- Raffreddamento a H2O nei nodi – generalmente tra 30° e 40°C
- Raffreddamento a H2O nei rack - generalmente tra 18° e 22°C
- Raffreddamento con liquido a due fasi - generalmente tra 30° e 40°C
- Raffreddamento immersivo – generalmente $\lesssim$ 30°C

MODUL'DATA CENTER
www.modul-data-center.com

7 cooling technologies used

	Advantages	Disadvantage	PUE mini Modul'Room*
Direct expansion	▯ Simple ▯ Low cost ▯ Low cost redundancy ▯ Compact ▯ Total intégration in Modul'Room	▯ Suitable for low and medium power (<10KW / rack) ▯ Not the best efficiency	Around 1,4
Direct expansion + Free cooling	▯ More efficient than direct expansion ▯ Compact	▯ Suitable for low and medium power (<10KW / rack) ▯ Not available for all models ▯ Not possible in every configuration ▯ Needs more maintenance	Around 1,25 depending on environmental conditions
Chilled water	▯ Suitable for large configurations and HPC ▯ Max Power = 35KW par rack ▯ "Hot power" scalability with our solution	▯ Cost ▯ Needs at least 2 chillers ▯ Chillers not integrated to shelters ▯ Less industrialized and requires work on site	Around 1,4
Chilled water + Free chilling	▯ Better efficiency than water cooling alone	▯ Idem water cooling ▯ ROI not guaranteed in hot environments	Around 1,2 depending on environmental conditions
Adiabatic	▯ Power efficiency ++ ▯ Fast ROI ▯ Reliability	▯ Less compact than direct expansion ▯ Not suitable for high humidity	Around 1,1 depending on environmental conditions
Direct Liquid Cooling (DLC)	▯ Energy Efficiency ++ ▯ No hard constraints ▯ Reliability	▯ Needs a water distribution ▯ Adapted to HPC	Around 1,1
Oil computing	▯ Power efficiency +++ ▯ Very fast ROI ▯ Future technology ▯ Adapted to harsh environmental conditions	▯ Non standard technology ▯ Specific operation ▯ Mainly suitable for HPC ▯ Specific servers	Around 1,03 and servers consumption 30% under

Raffreddamento free-cooling diretto – stima PUE



Dato il grafico delle temperature medie nell'arco del 2024 in una citta' come Bologna, assumiamo che la temperatura dell'aria superi 18°C nei mesi di

Maggio, Giugno, Luglio, Agosto, Settembre, Ottobre.

Semplificando il calcolo, assumiamo che per temperature dell'aria esterna inferiori a 18° non sia necessario alcun raffredamento mentre per temperature superiori a 18°C si debba raffreddare l'aria.
Semplificando ulteriormente assumiamo che se non si raffredda PUE=1, mentre se si raffredda PUE=1,4. Facendo a questo punto una media pesata sui giorni "freddi" rispetto a quelli "caldi" otteniamo

PUE medio raffreddamento a aria = (180*1 + 185*1,4)/365 = 1,2

Stime di PUE a Bologna per un sistema raffreddato DLC + AIR in free-cooling

Let o(x) be the overhead of cooling for x = warm water (w), cold aisle air (c) and machine room air (a), then the PUE of a direct water-cooled system can be described as:

$$PUE = p * o(w) + q * o(c) + r * o(a).$$

With $p + q + r = 1$, the respective ratios of heat cooled by warm water, the RDHx portion (r) is dissipated into the machine room air. For a system with RDHx we can assume $r = 0$ – no heat will be dissipated from the rack into the machine room air. If we further assume the machine room cold aisle temperature to be close to the temperature of the warm water loop (this is subject to shared machine room optimization). We can assume that using a free cooling solution to cool air and water and considering an inlet water temperature at 40°C degrees as similar as air on cool aisle we can use the following overheads:

$$o(w) = 1.05 \text{ and } o(c) = 1.2$$

With $p = 85\%$ and $q = 1 – p = 15\%$ we get:

$$PUE = 0.85 * 1.05 + 0.15 * 1.2 = 1.07$$

Power calculation of the node gives us a power consumption of about 2200 W under HPL like load. This leads to a rack power consumption of about 80KW per rack with 36 compute nodes. For the above given span of nodes 3120 we calculate a power range between:

5.9 MW to 7.2MW under HPL.

Experience shows that the average power consumption of the system will be less than 70% of HPL.

Lo standard *Open Compute Project* per raffreddamento a liquido

Il settore Hyperscaler/Cloud ha stabilito alcune specifiche comuni per i server e i rack raffreddati a acqua per una maggiore flessibilita' dell'infrastruttura (sostituzione server senza cambiare l'infrastruttura) e costi ridotti

Rafrreddamnto a acqua in OCP comporta generalmente ∽ 80% acqua e ∽ 20% aria

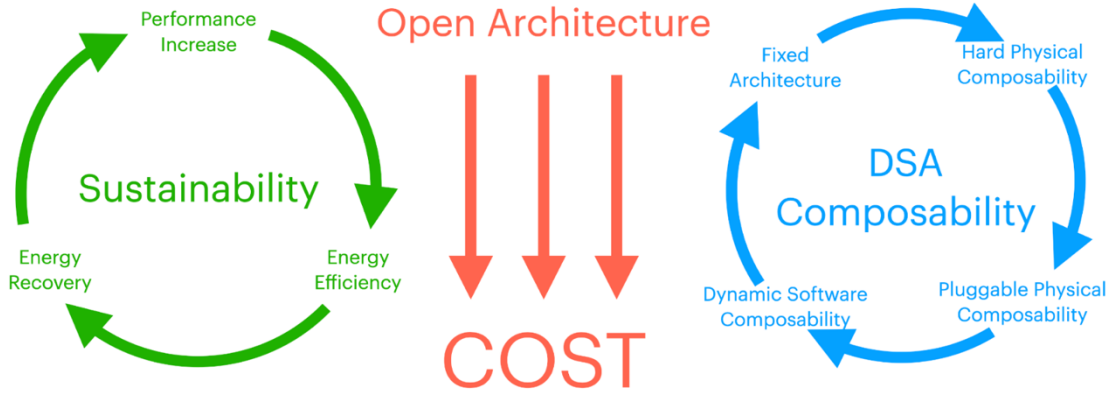


Figure 2 - Core areas that are driving Computer Architecture Today

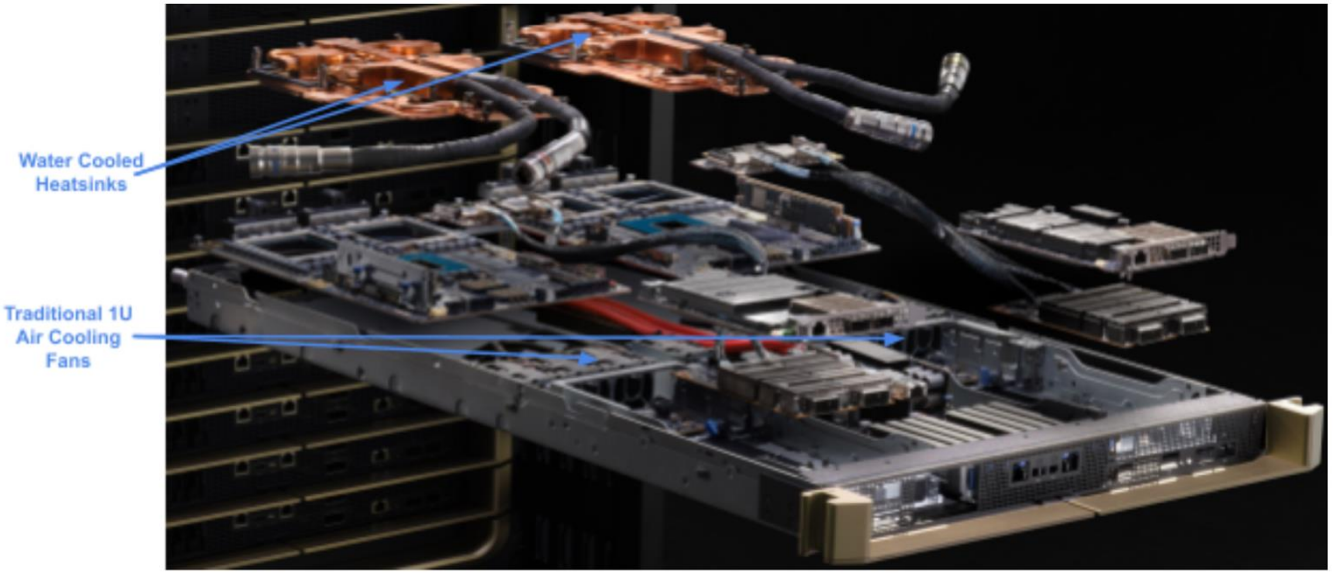
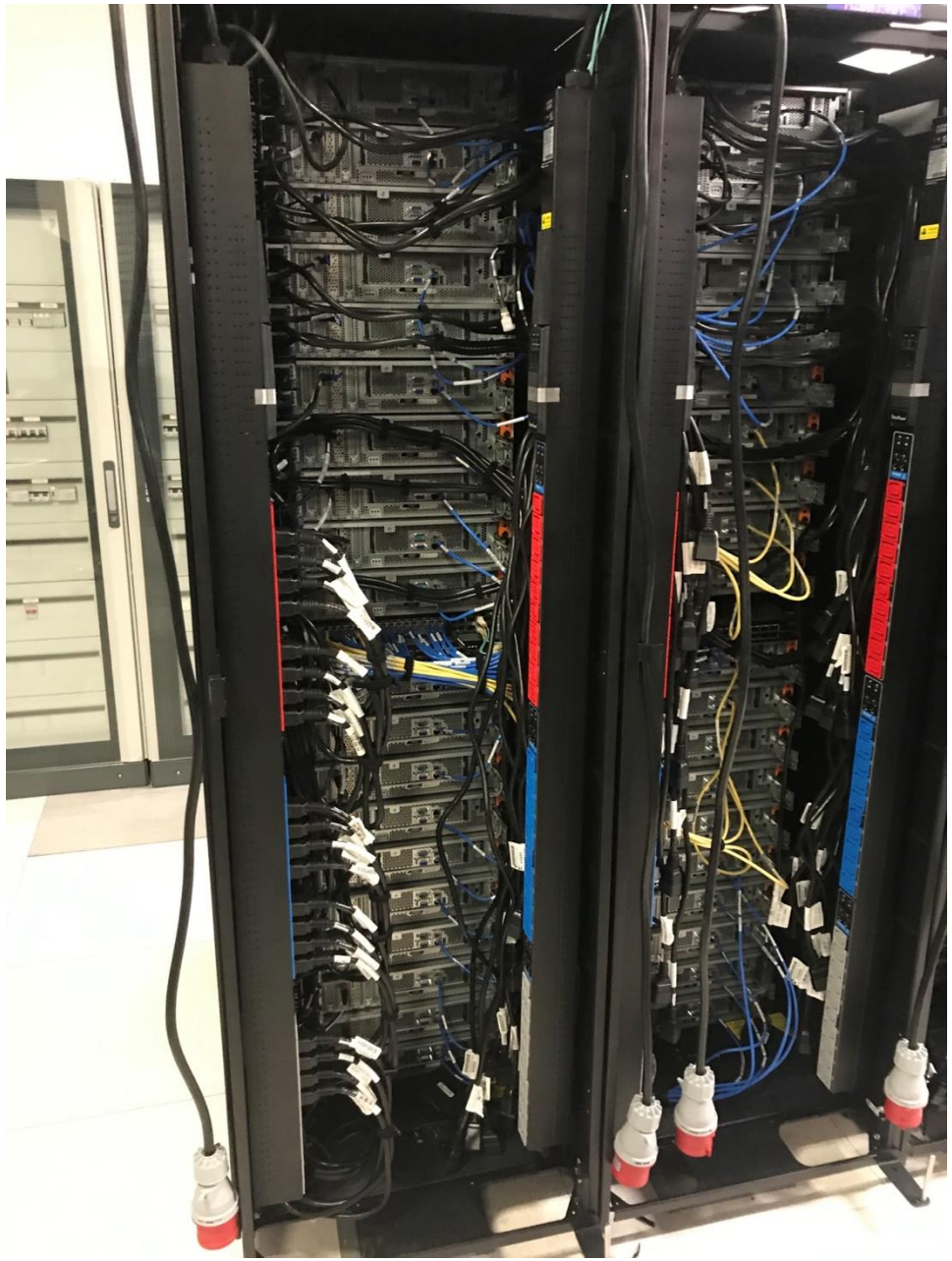


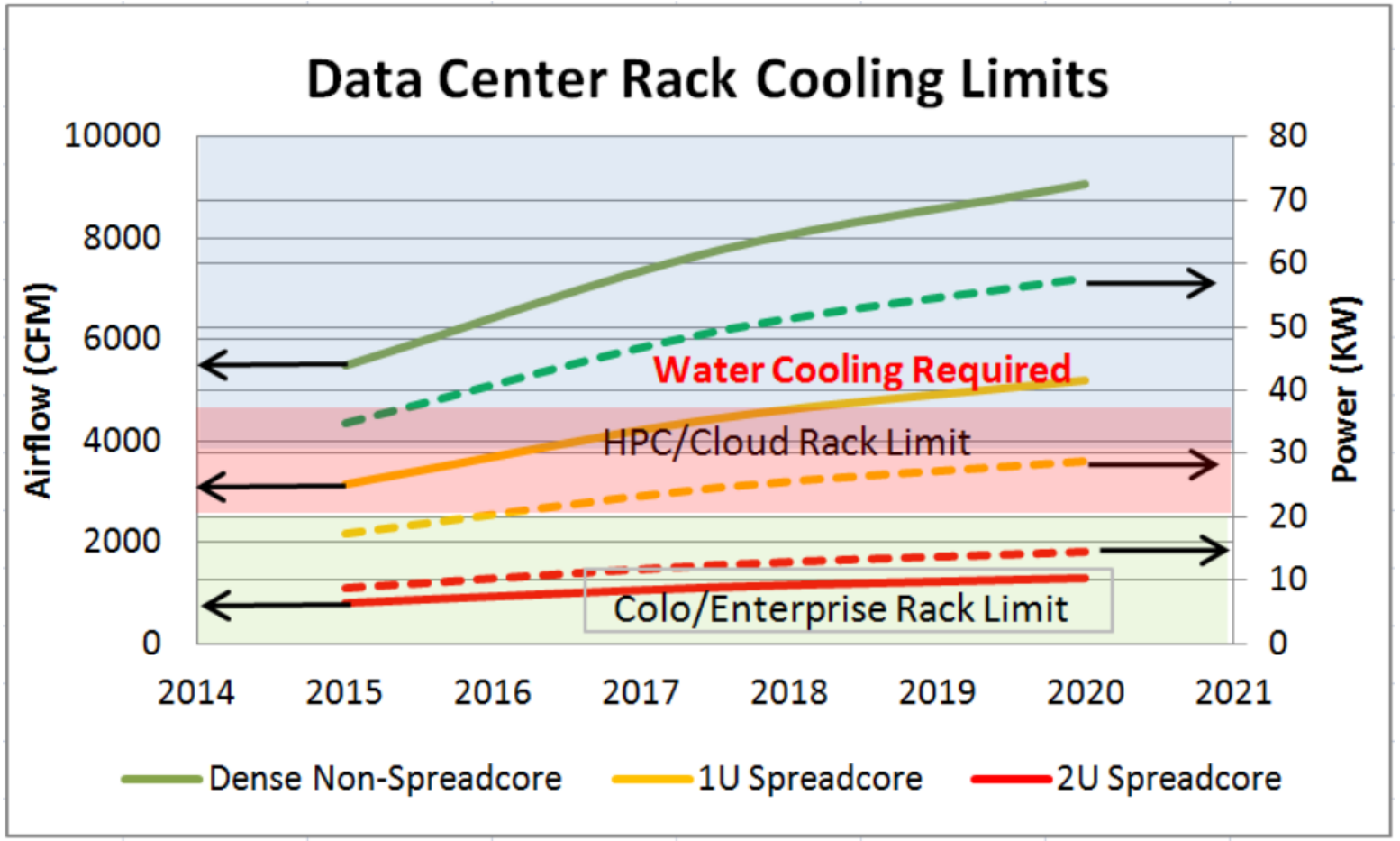
Figure A3 - NVidia Hybrid Cooled Grace Blackwell Chassis

Un rack ha varie componenti che possono influenzare il corretto raffreddamento del sistema



Data Center Level Cooling and Power Limits

- Node power density trends cannot be cooled at data center level
  - Partial rack population or rack level power capping may be required for Dense and 1U



PUE, ITUE and ERE

- PUE

$$PUE = \frac{\text{Total Facility Power}}{\text{IT Equipment Power}}$$

  - **Power usage effectiveness** (**PUE**) is a measure of how efficiently a computer data center uses its power;
  - PUE is the ratio of total power used by a computer facility[] to the power delivered to computing equipment.
  - Ideal value is 1.0
  - It does not take into account how IT power can be optimised

- ITUE

$$ITUE = \frac{(\text{IT power + VR + PSU + Fan})}{\text{IT Power}}$$

  - **IT power effectiveness** (ITUE) measures how the node power can be optimised
  - Ide

- ERE

$$ERE = \frac{\text{Total Facility Power} - \text{Treuse}}{\text{IT Equipment Power}}$$

  - **Energy Reuse Effectiveness** measures how efficient a data center reuses the power dissipated by the computer
  - ERE is the ratio of total amount of power used by a computer facility[] to the power delivered to computing equipment.
  - An ideal ERE is 0.0. If no reuse, ERE = PUE

PUE, ITUE and ERE

- PUE

$$PUE = \frac{\text{Total Facility Power}}{\text{IT Equipment Power}}$$

  - **Power usage effectiveness** (**PUE**) is a measure of how efficiently a computer data center uses its power;
  - PUE is the ratio of total power used by a computer facility[] to the power delivered to computing equipment.
  - Ideal value is 1.0
  - It does not take into account how IT power can be optimised

- ITUE

$$ITUE = \frac{(\text{IT power} + \text{VR} + \text{PSU} + \text{Fan})}{\text{IT Power}}$$

  - **IT power effectiveness** (ITUE) measures how the node power can be optimised
  - Ideal value if 1.0

- ERE

$$ERE = \frac{\text{Total Facility Power} - \text{Treuse}}{\text{IT Equipment Power}}$$

  - **Energy Reuse Effectiveness** measures how efficient a data center reuses the power dissipated by the computer
  - ERE is the ratio of total amount of power used by a computer facility[] to the power delivered to computing equipment.
  - An ideal ERE is 0.0. If no reuse, ERE = PUE

Considerazioni preliminari sul raffreddamento di sistemi a alta densita' computazionale per rack

Rack >= 35kW IT non possono essere raffreddati efficacemente con aria anche a temp <=25°C

L'energia spesa nel raffreddamento a aria e' spesso oltre il 30% dell'energia IT

Sistemi ibridi per raffreddamento aria-acqua (aria fredda front e pannelli refrigerati a acqua retro) possono consentire di migliorare l'efficienza ma in un limite di rack >=40kW IT

Sistemi con DLC parziale (CPU e GPU) possono estrarre fino a 75% calore in acqua

Ulteriore sviluppo per soluzioni con DLC parziale e pannelli retro raffreddati a acqua estraggono quasi 100% in acqua ma richiedono comunque aria fredda (<=32°C) sulla parte frontale

Raffreddamento a acqua DLC completo e' la soluzione per sistemi per rack>=70kW IT

Calore specifico dell'aria: $c_{p,a}$=1 kJ/kg$_a$K

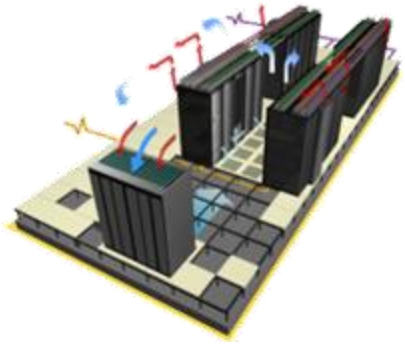
Calore specifico dell'acqua: $c_{p,ac}$=4,81 kJ/kg$_v$K

Quantita' di energia per riscaldare o raffreddare di 1°C un Kg

Rapporto calore specifico H_2O/A ~ 5

Cooling comparison

Air Cooled

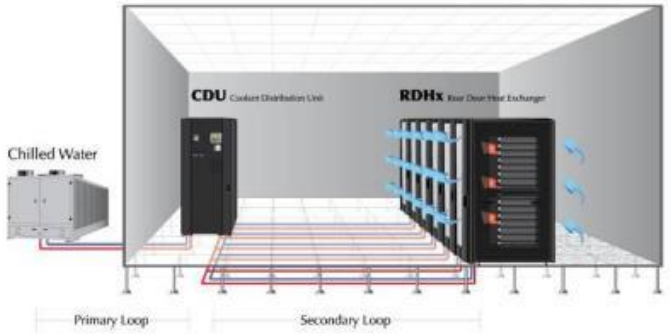


- Standard air flow with internal fans
- Fits in any datacenter
- Maximum flexibility
- Broadest choice of configurable options supported
- Supports Native Expansion nodes (Storage NeX, PCI NeX)

PUE ~2 – 1.5

Choose for broadest choice of customizable options

Air Cooled with Rear Door Heat Exchangers



- Air cool, supplemented with RDHX door on rack
- Uses chilled water with economizer (18°C water)
- Enables extremely tight rack placement

PUE ~1.4 – 1.2

ERE ~ 1.4 – 1.2

Choose for balance between configuration flexibility and energy efficiency

Direct Water Cooled

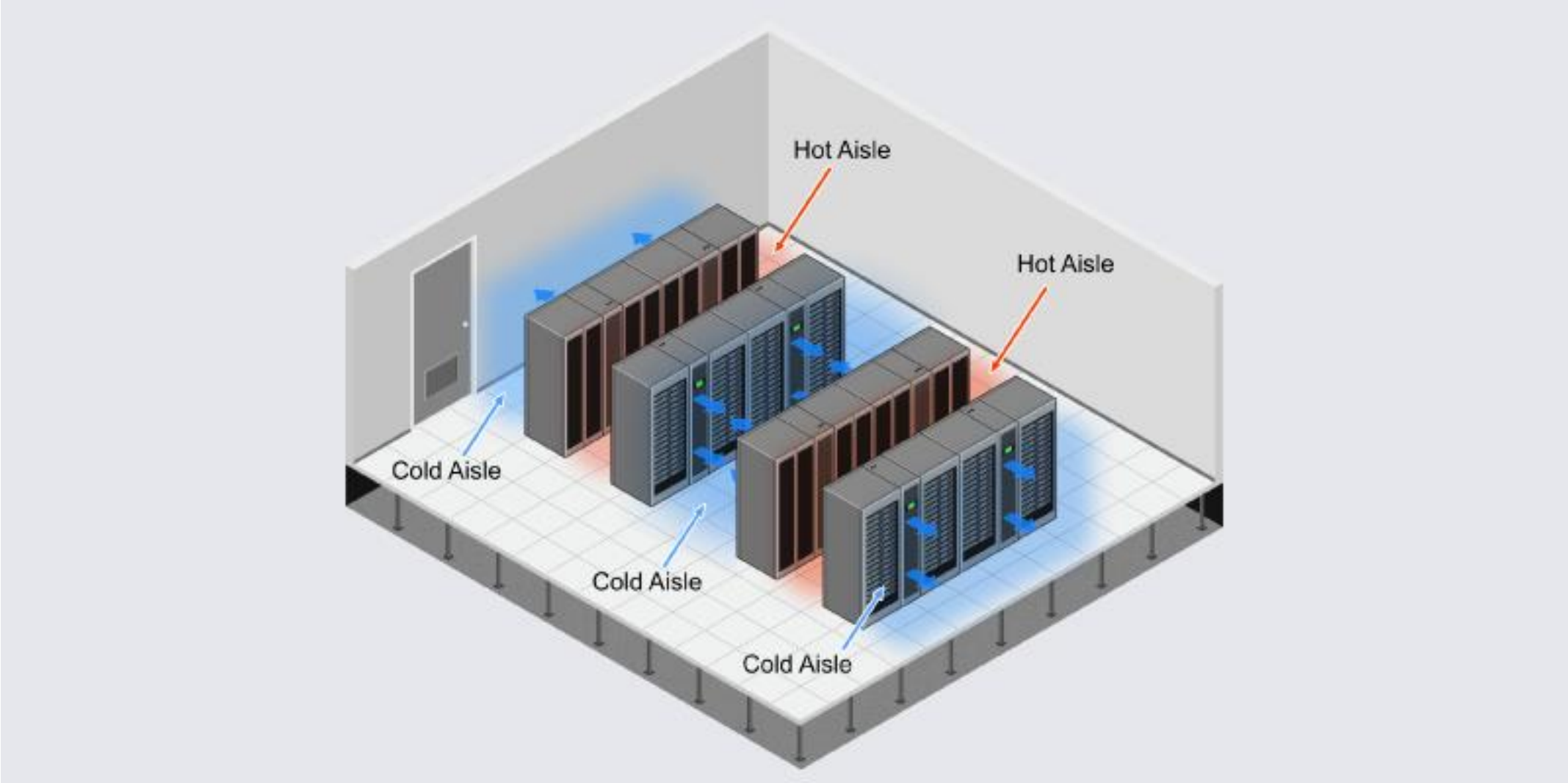
- Direct water cooling with no internal fans
- Higher performance per watt
- Free cooling (45°C water)
- **Energy re-use**
- Densest footprint
- Ideal for geos with high electricity costs and new data centers
- Supports highest wattage processors

PUE ~ 1.1

ERE < < 1 with hot water

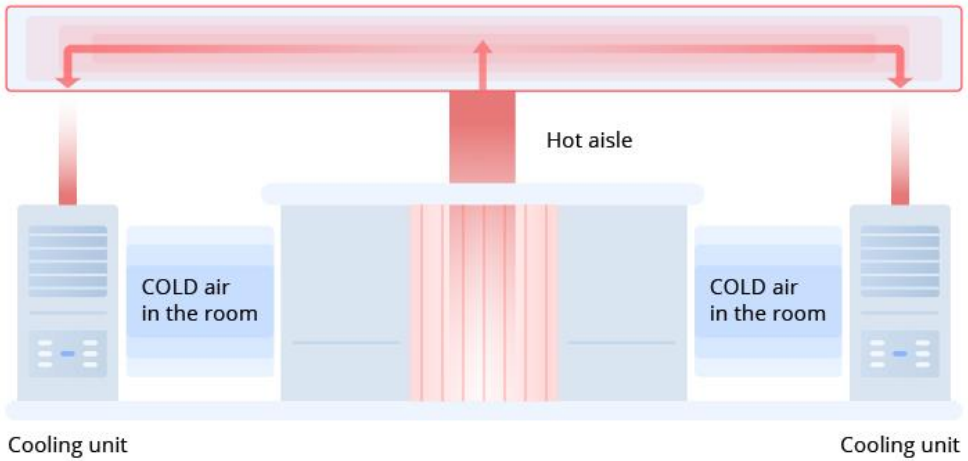
Choose for highest performance and energy efficiency

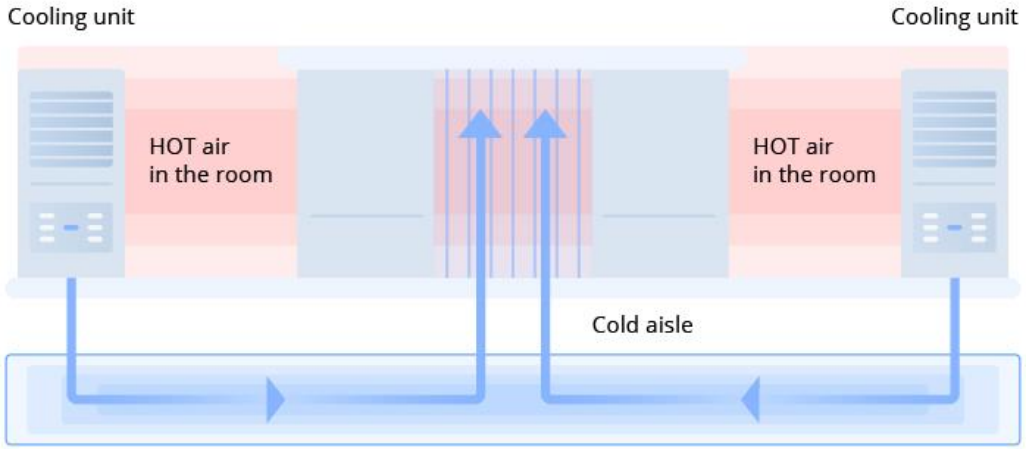
Raffreddamento a aria: generalmente non particolarmente efficiente - PUE >= 1,3



https://www.vertiv.com/it-emea/solutions/learn-about/liquid-cooling-options-for-data-centers/

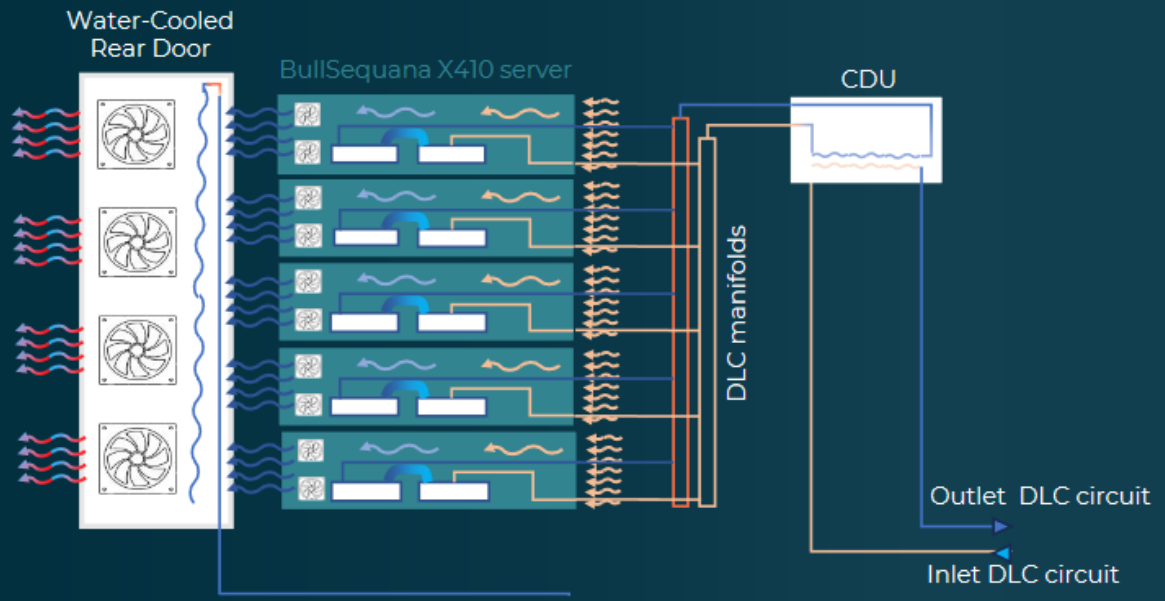
Raffreddamento a aria: circolazione aria fredda e aria calda possibilmente compartimentati





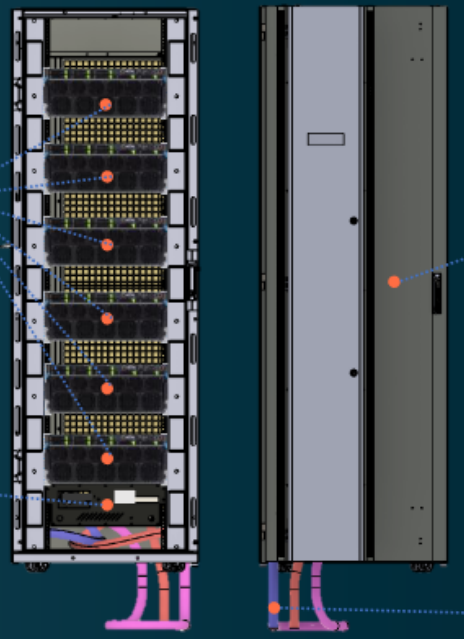
https://community.fs.com/it/article/data-center-containment-types-benefits-and-challenges.html

2nd gen BullSequana X400 DLC Platform



6x 5U accelerated compute server 8-GPU

AMD NVIDIA



1x RDHx

1x CDU

1x inlet water supply for both CDU and RDHx

Front view

Rear view

EVIDEN

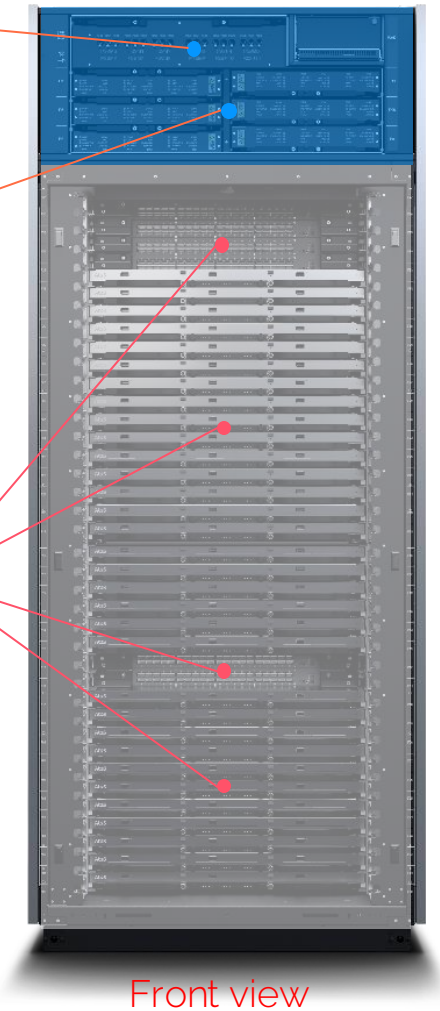
Infrastructure architecture overview

1x Front Compute PDU

Power Group

6x Front DLC PSU Shelves

DLC ready Compute Rack



Front view

Rear view

1x Hydraulic PDU

1x Rear Compute PDU

6x Rear DLC PSU Shelves

1x Expansion Tank

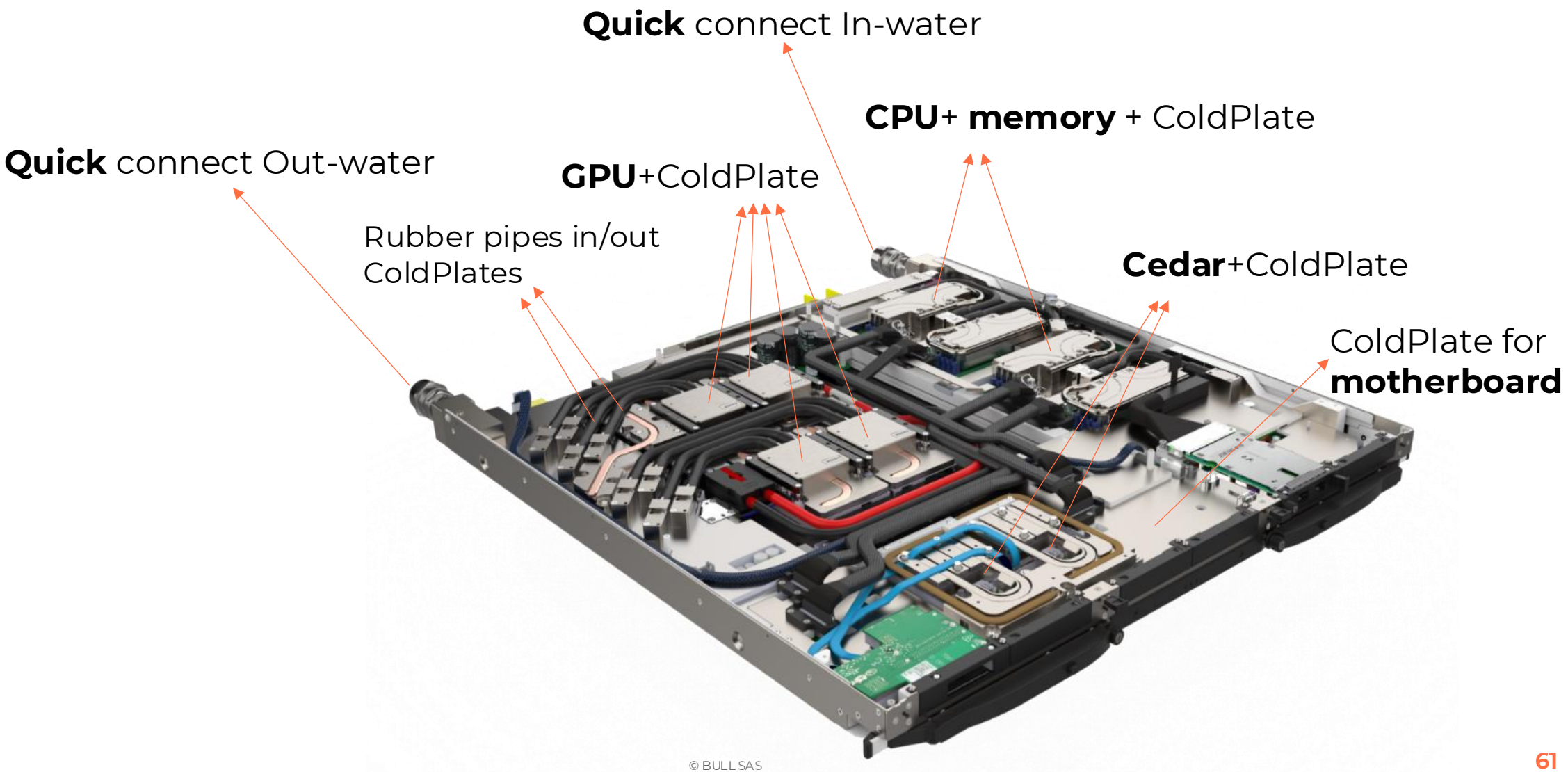
2x Management switches

2x Hydraulic Pumps

Hydraulic Group

1x Heat Exchanger

WATERCOOLED COMPONENTS IN BULLSEQUANA X3145H DLC BLADE

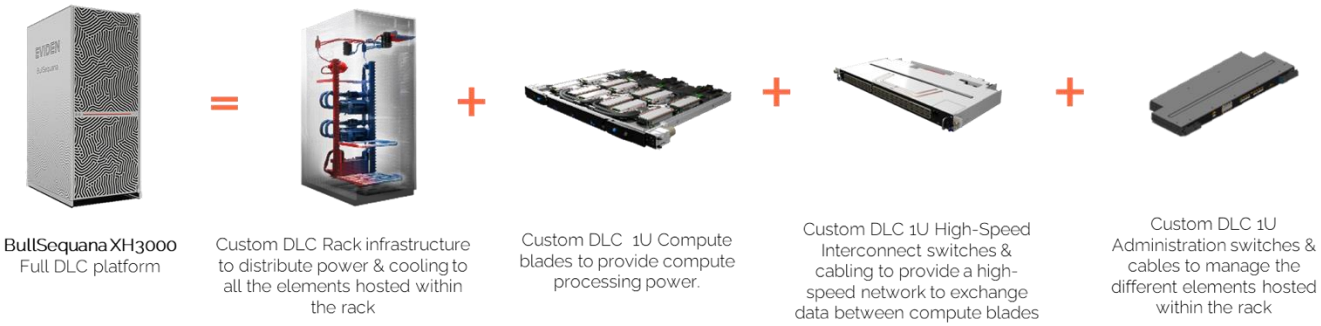


WATER COOLING TEMPERATURE & FLOWRATES

Scenario	Inlet water temperature	Outlet water temperature	Flow rate
1	32°C	41.9°C	228 L/Min
2	33°C	43°C	227 L/Min
3	34°C	44°C	227 L/Min
4	35°C	45°C	226 L/Min

DIRECT LIQUID COOLING DESIGN

Over 97% heat capture through DLC technology



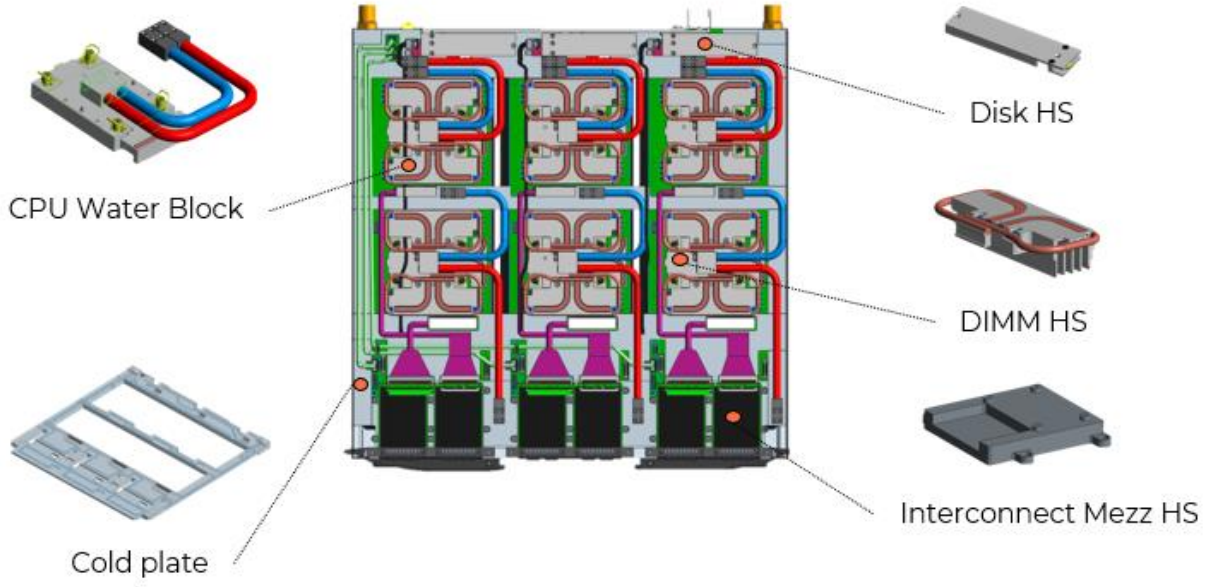
BullSequana XH3000
Full DLC platform

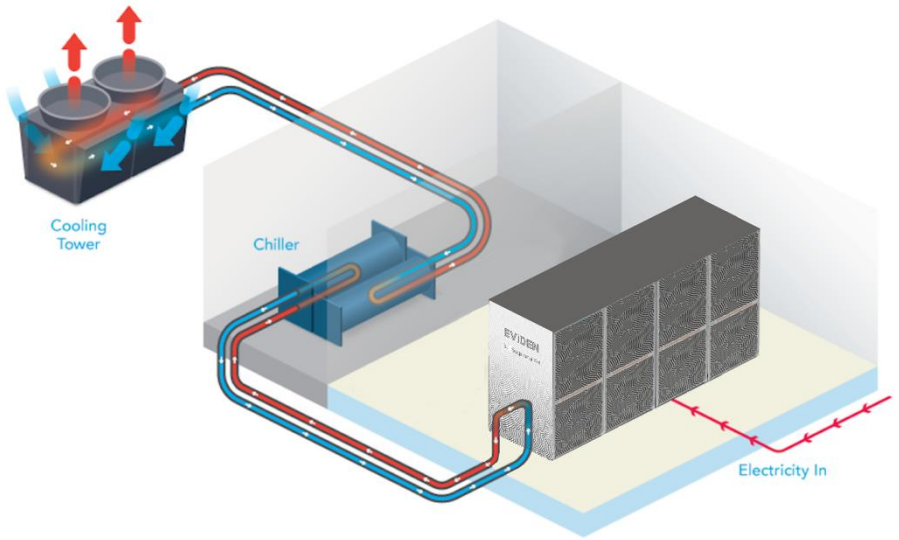
Custom DLC Rack infrastructure to distribute power & cooling to all the elements hosted within the rack

Custom DLC 1U Compute blades to provide compute processing power.

Custom DLC 1U High-Speed Interconnect switches & cabling to provide a high-speed network to exchange data between compute blades

Custom DLC 1U Administration switches & cables to manage the different elements hosted within the rack

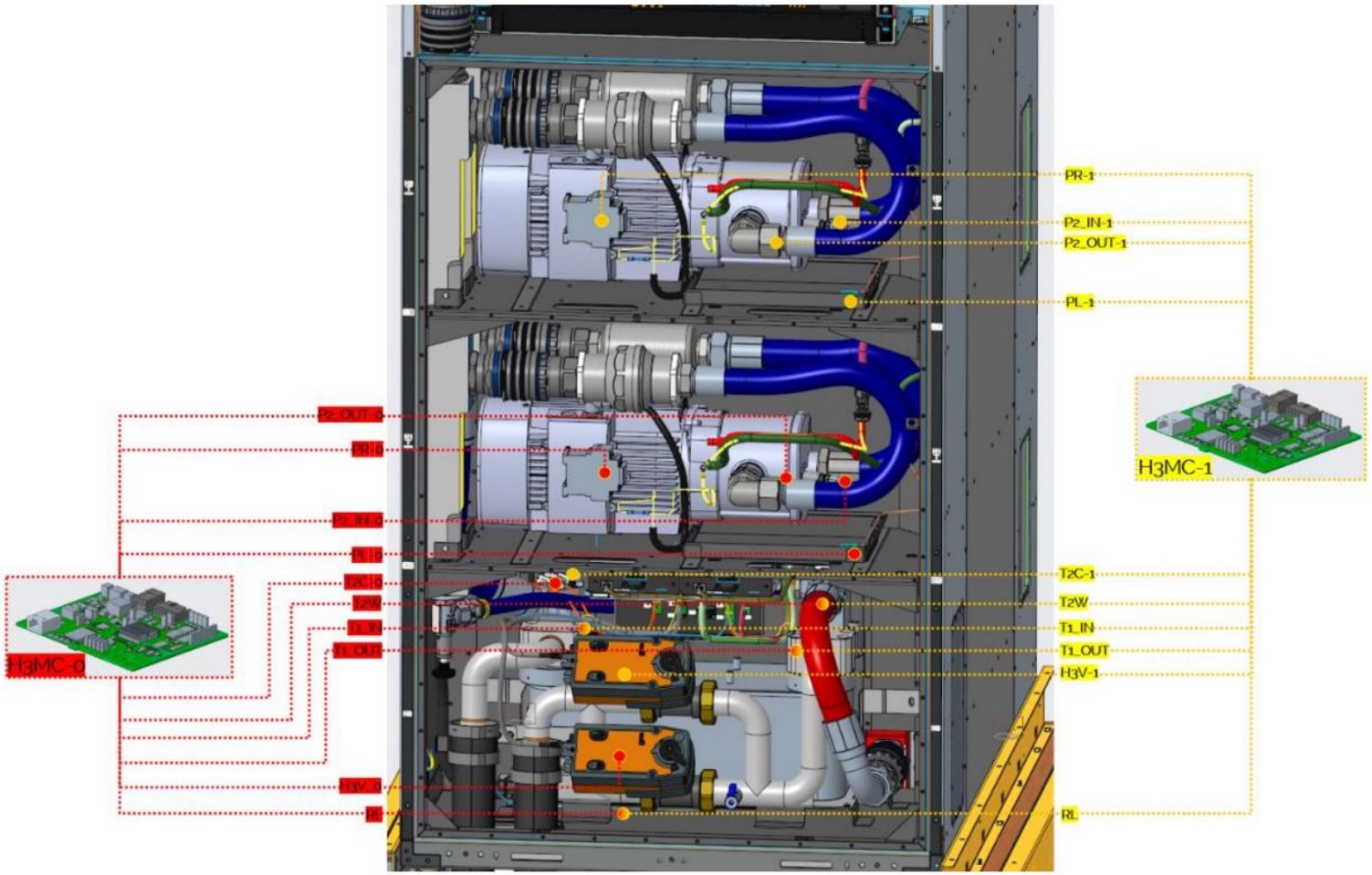


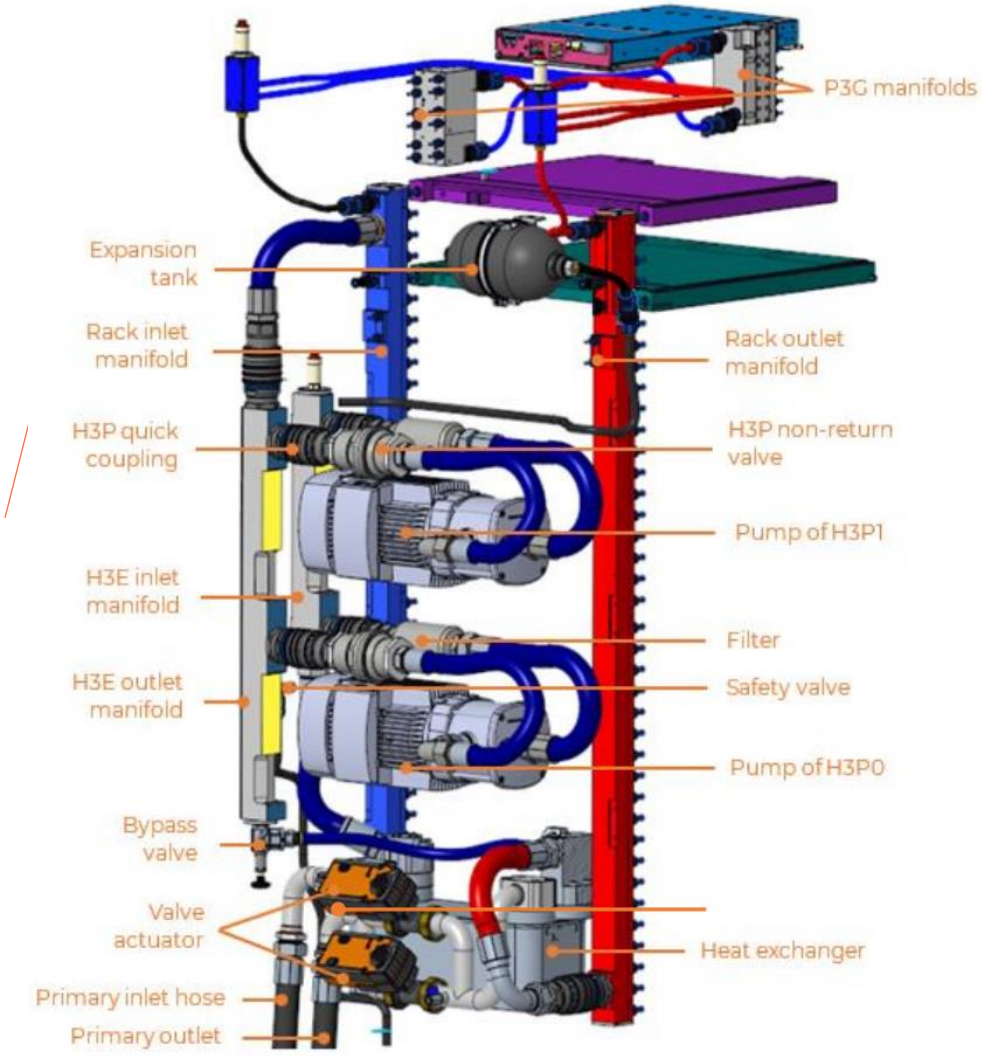


Higher rack inlet water temperature → **Less use of chillers** → **Energy bill savings** → **Lower TCO** → **Higher rack inlet water temperature**

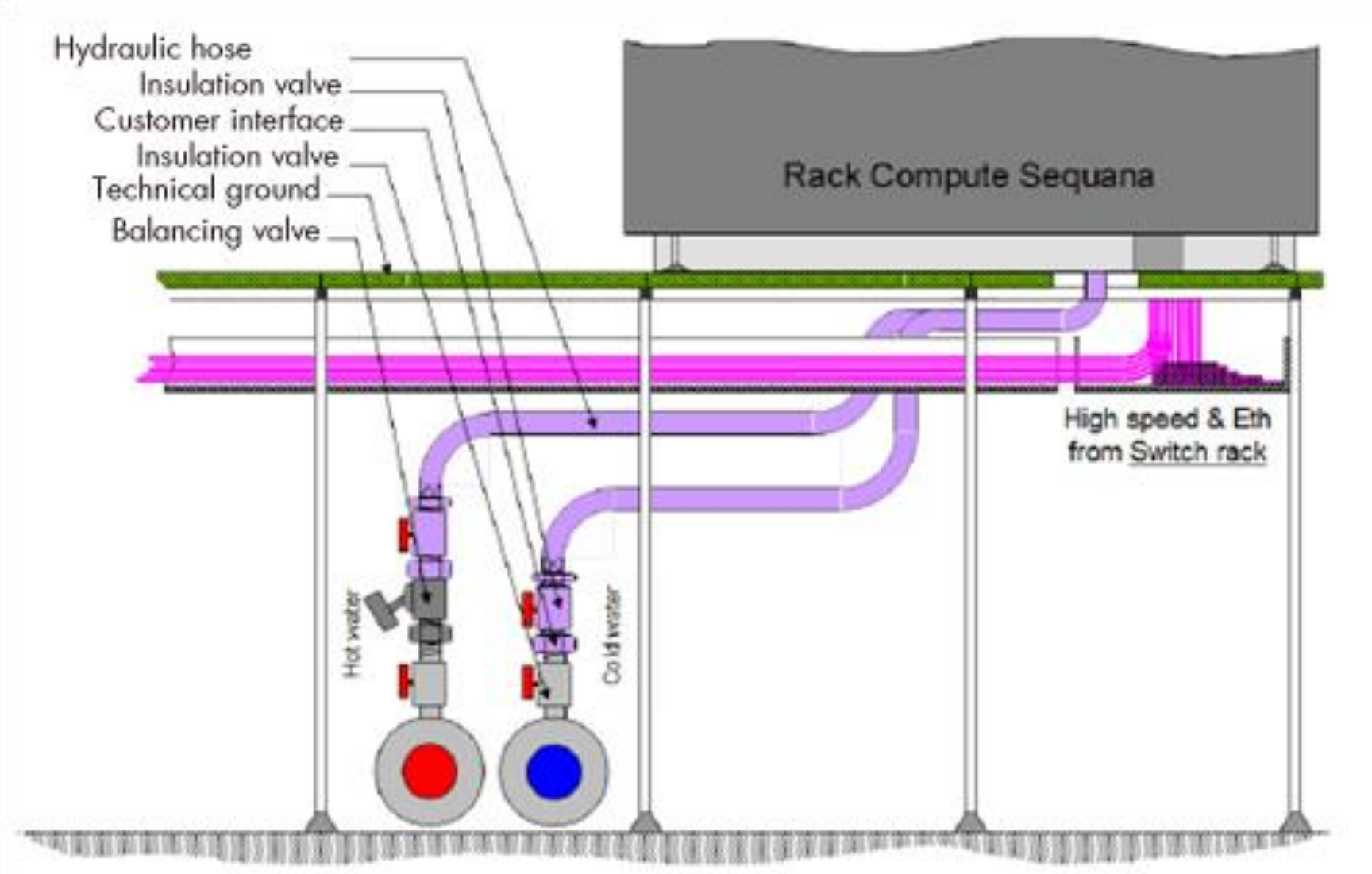
EVIDEN

WATERCOOLED COMPONENTS IN BULLSEQUANA X3145H DLC BLADE



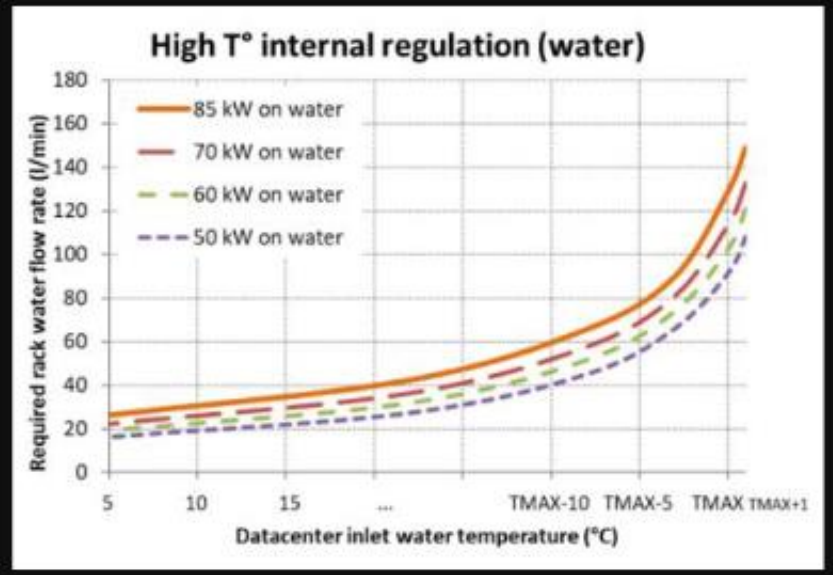


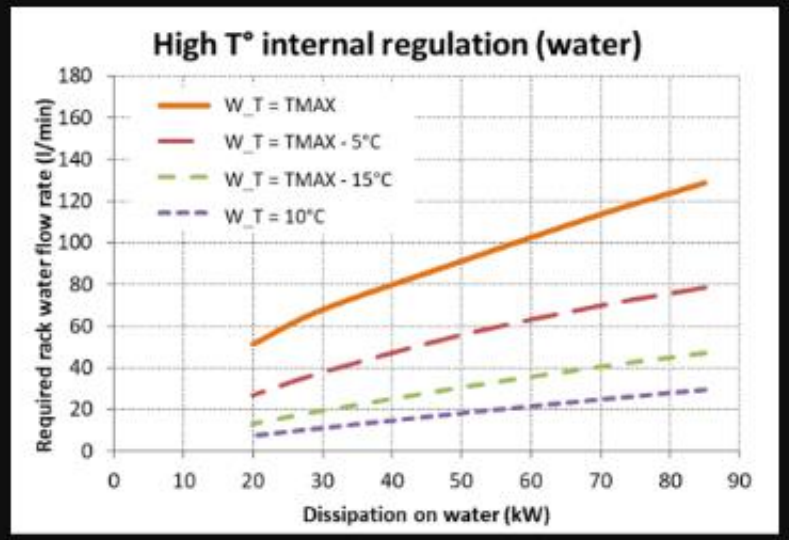
XH3000 RACK CONNECTION TO CUSTOMER WATER CIRCUIT

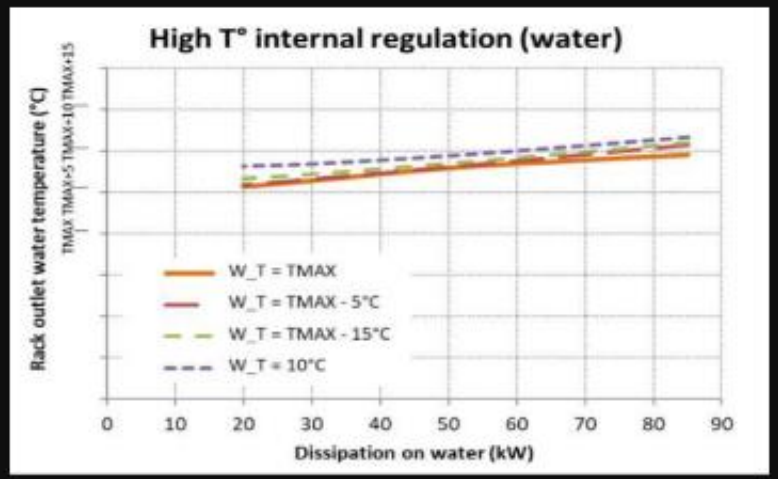


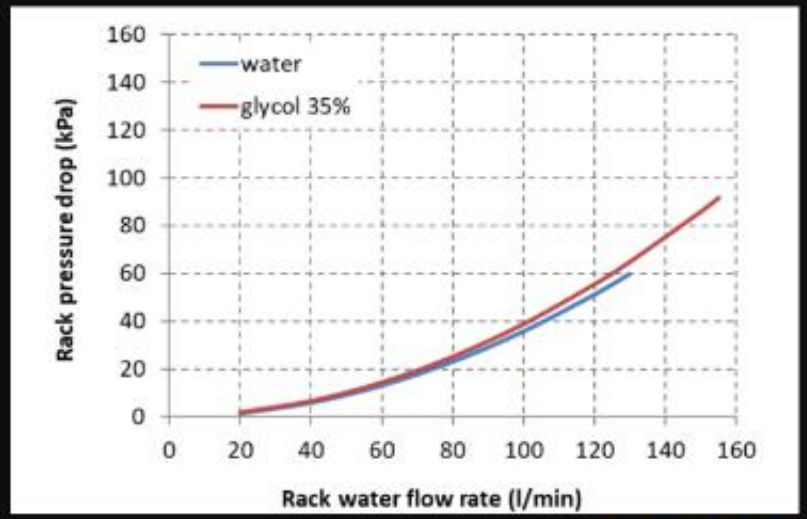
Power & Cooling
Booster Partition

Booster partition
- TMAX = 36°C
- Active / Active HYC setup

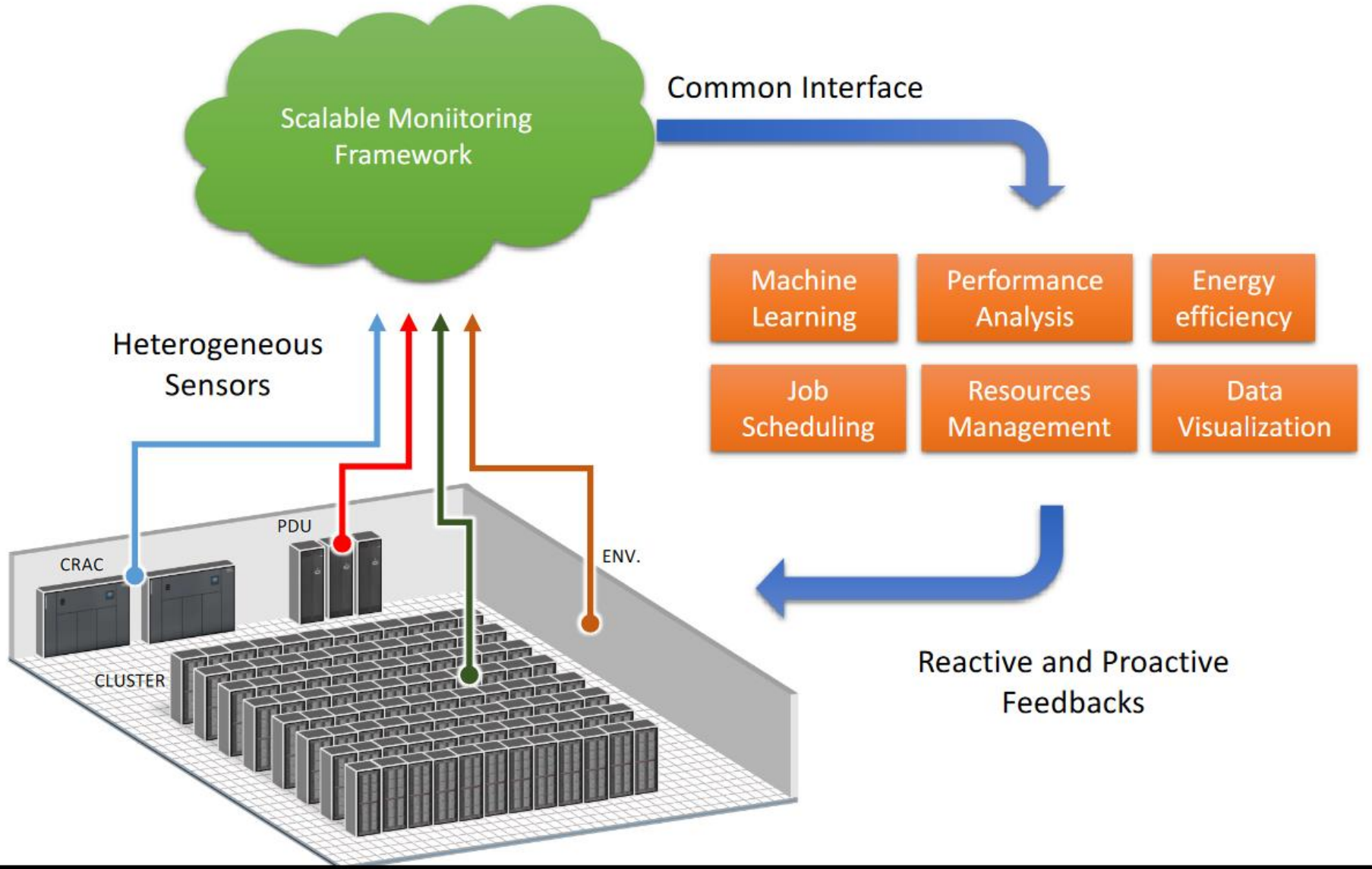








A New Trend: Datacentre Automation



Courtesy by Andrea Bartolini – DEI UniBO

Formula per il Total Cost of Ownerhip (TCO) nelle gare Europee per i sistemi pre e Exascale

3.2.2 TCO model

To evaluate the second factor of the scoring metric descripted in Section 3.4, CINECA considers a TCO methodology with the following approximated equation:

$$TCO = CAPEX_{(aq)} + OPEX_{(en)}$$

For the current procurement CINECA will assume:

$CAPEX_{(aq)}$	Acquisition cost	The value reported in the tender
$OPEX_{(en)}$	Energy cost	Evaluated through the benchmark procedure, see the following.

Table 3-3: List of CAPEX and OPEX definition

Therefore, to evaluate the impact of the offered solution on the Total Cost of Ownership, CINECA will consider for the current procurement only the operational energy cost ($OPEX_{(en)}$) of the TCO.

For each application App of the benchmark suite an estimation of the energy consumed by a single application for the entire lifetime of the system is defined as follows:

$$Energy.LifeTime.App = PUE * n.App * ETS.App$$

PUE_{DLC}	Power Usage Effectiveness for a DLC solution	1.08[1]
PUE_{AC}	Power Usage Effectiveness for an AC solution	1.35
LifeTime	LifeTime of the system	157.680.000 seconds
Price.Energy	In Euro/kWh	0.15

Table 3-4: Values to provide to calculate the TCO.Energy

Assuming:

- A Booster (B) partition under direct liquid cooling (DLC);
- A portion *x* of the Data Centric and General Purpose (DCGP) partition under air cooling (AC) and the rest (1-*x*) under DLC;

The computation of Energy.LifeTime.App becomes:

$$Energy.LifeTime.App(B) = PUE_{DLC} * n.App(B) * ETS.App(B)$$

$$Energy.LifeTime.App(DCGP) = (1-x)PUE_{DLC} * n.App(DCGP) * ETS.App(DCGP) + x * PUE_{AC} * n.App(DCGP) * ETS.App(DCGP)$$

Where:

$$n.App(B) = \frac{Nodes.Total(B)}{Nodes.App(B)} * \frac{LifeTime}{TTS.App(B)}$$

$$n.App(DCGP) = \frac{Nodes.Total(DCGP)}{Nodes.App(DCGP)} * \frac{LifeTime}{TTS.App(DCGP)}$$

The terms refer to the following meaning:

- Energy.LifeTime.App is the energy consumed if the system were to run only the application App for the entire LifeTime of the system;
- PUE is power usage effectiveness of the datacentre and will be provided to the Candidates;
- Nodes.Total number of nodes in the offered system solution;
- Nodes.App are the nodes used to execute the benchmark of the application App;
- LifeTime is the existence time of the system;
- TTS.App is the elapsed time to execute the benchmark of the application App;
- ETS.App is the value of energy consumed to execute the benchmark of the application App.

The TCO energy component of an application App can be calculated as follows:

$$TCO.Energy.App = [Energy.LifeTime.App(B) + Energy.LifeTime.App(DCGP)] * Price.Energy$$

Where Price.Energy is the energy price that will be assumed to be constant for the entire LifeTime of the system. The TCO.Energy will be estimated as the average value:

$$TCO.Energy = \frac{\sum_{App} TCO.Energy.App}{N}$$

Where N is the number of applications of the benchmark suite. The Candidate is asked to provide ETS, TTS and Nodes.App for each component of the benchmark suite as described in the in the document "LOT3-Benchmark_Information". To calculate TCO.Energy, CINECA will provide the following values:

Stima del Total Cost of Ownerhip (TCO) per Leonardo - Cineca

	QE	Pluto	SpecFEM3D	MilC
$Nodes.App(B)$	12	36	16	12
$TTS.App(B)$	435	3214.4	275	175
$ETS.App(B)$	1.011	10.95	1.680	0.45
$n.App(B)$	104,395,034	4,709,209	123,850,473	259,496,229
$Energy.LifeTime.App(B)$	114,936,741	56,155,193	226,586,917	127,166,127
$Nodes.App(DCGP)$	14	18	40	14
$TTS.App$(DGCP)	992.9	2051.9	366.4	2313.6
$ETS.App(DCGP)$	2.14	4.82	2.14	4.23
$n.App(DCGP)$	17,423,455	6,557,513	16,525,415	7,477,416
$Energy.LifeTime.App(DCGP)$	40,604,665	34,420,252	38,511,818	34,444,490
$TCO.Energy.App$	€ 23,331,211	€ 13,586,317	€ 39,764,810	€ 24,241,593
$TCO.Energy$	€ 25,230,983			

Summary and comments – 5th & 6th Lessons

- Storage architecture and its own scalability

- Block and object storage

- Filesystem

- IO benchmark

- Data center and cooling solutions

- Energy efficiency

- Workload distribution and cooling optimization

Thank You

Marco Briscolini, PhD

marco.briscolini@gmail.com
Cell: 3357693820

03/19/2025