Data and Performance

Summer School
Sept. 2025

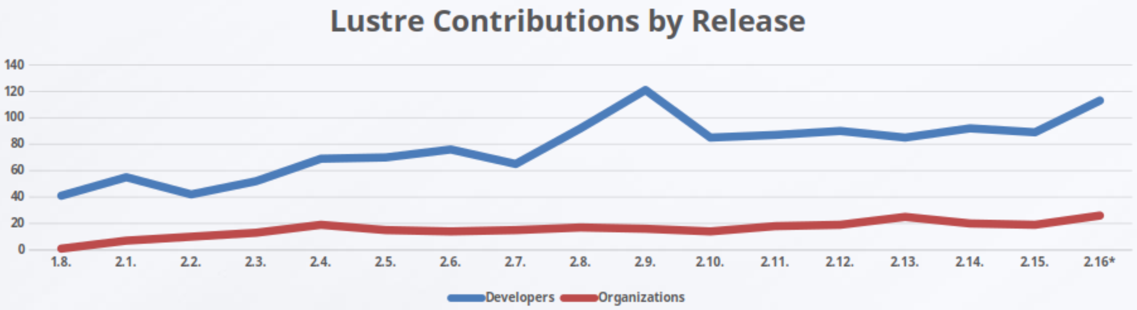
Jean-Thomas Acquaviva
jtacquaviva@ddn.com

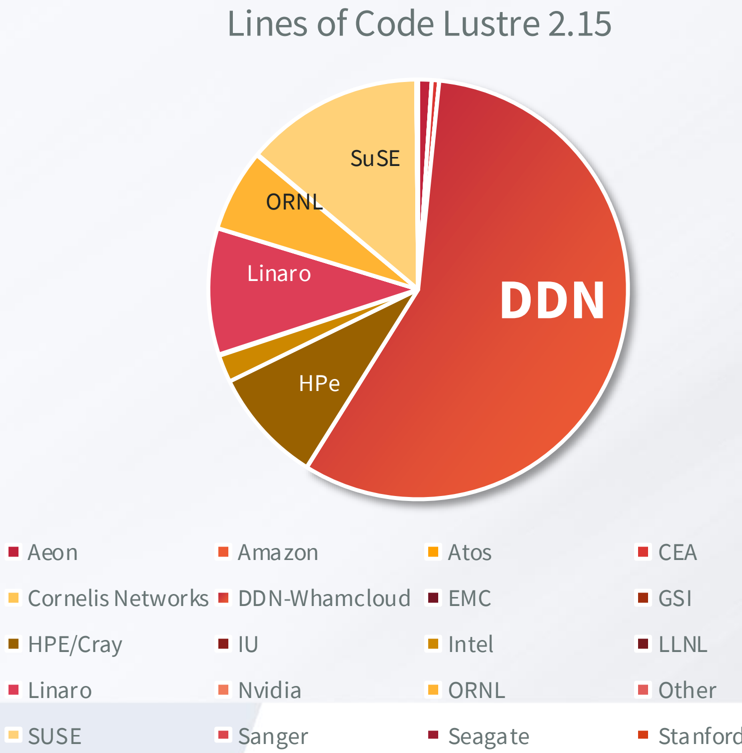
Who are we?

DDN an Open-Source Driven Company

THE AI DATA COMPANY

Lustre Open-Source Parallel FRile system (OpenSFS)





- Designed for HPC: data extension of the compute platform
- OpenSFS provides overall directions and a forum for discussion among users
- DDN is the lead contributor to Lustre
- User meetings in Europe organized by EOFS
- User meetings in Asia organized by DDN

EOS NVIDIA Flagship System

576 DXH: 4608 H100 GPU

NDR400 IB Compute and Storage

Storage:

48 AI400NVX2

EXAScaler 6

12 PB flash

4.3 TB/s Read

3.1 TB/s Write

DDN **Hot node** for Accelerated AI Training

EOS IS THE THIRD-GENERATION FLASGHIP DGX SUPERPOD

HPC for science: AI a

Centre National de la Recherche Scientifique

Jean Zay - World Class AI system

a BigScience initiative

BLOOM

176B params · 59 languages · Open-access

- 1000,000 GPU hours of learning

- **IDRIS is driving excellence in scientific research for modeling and intensive numerical computation which requires seamless scaling and enterprise resilience.**
- Supplied as a service infrastructure, this is next generation class systems used for the refinement of AI algorithms at large scale.

DDN European Collaboration Framework

A – EuroHPC Infrastructures powered by DDN

Leonardo


Meluxina

Discoverer

Vega


Deucalion

B – EuroHPC Research Programs and DDN

- EuroHPC design of next-Gen IO system
- AI-automated features extractions from Satellite Images
- Nuclear Fusion code optimization

C – DDN R&D Spending in Europe

- Significant portion of our WW turnover is spent on R&D
- 25 persons R&D in EU
- France focused on Software Platforms

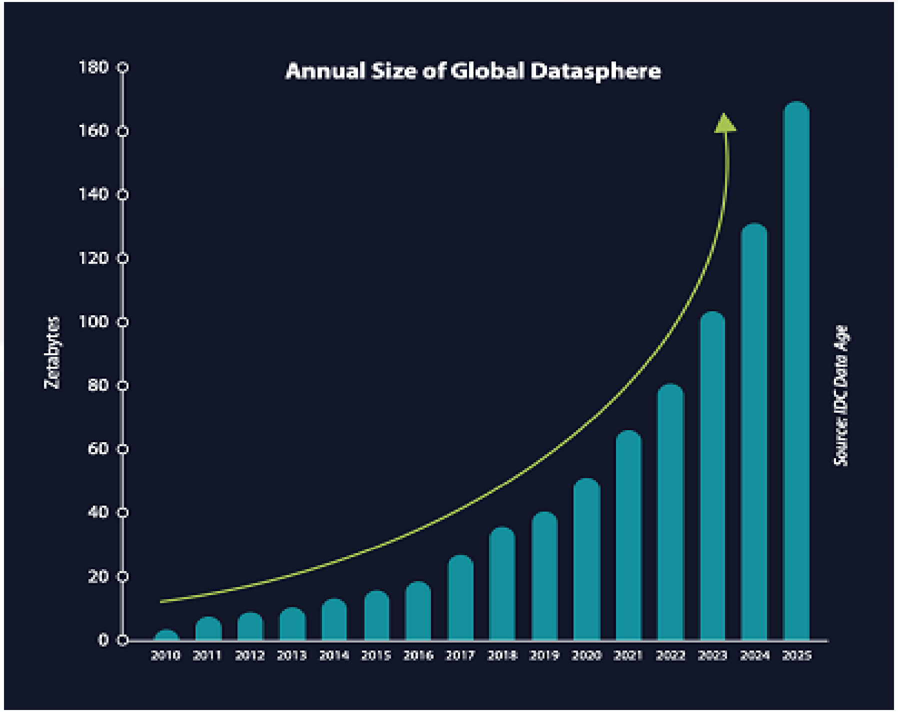
Why do we matter?

Understanding Data at Scale: Commodity and Scarcity

- 2025: Data acquisition devices are ubiquitous
- 2025: Computing capabilities are ubiquitous

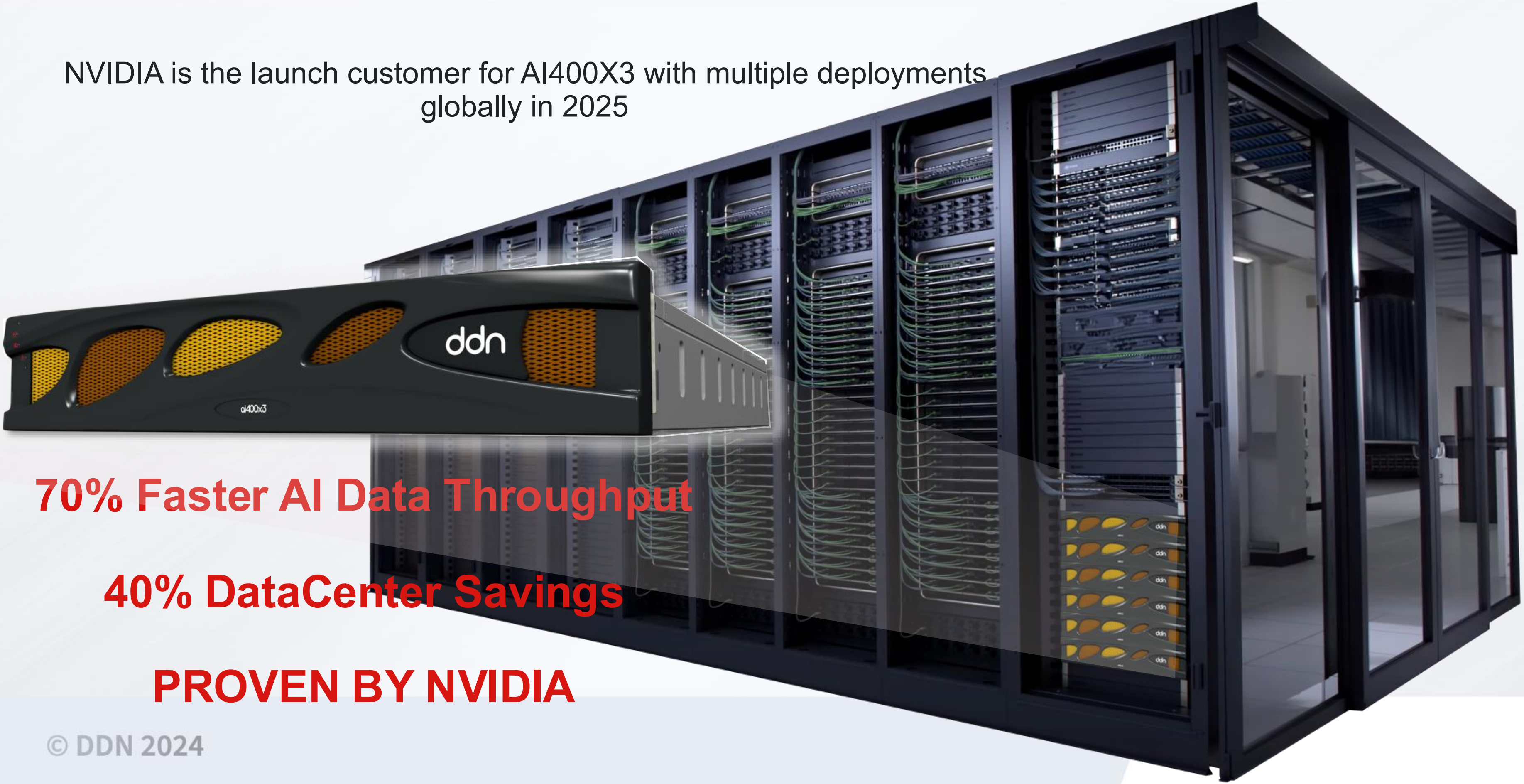
- 2025+: Data management at scale is scarce

As computing is becoming a commodity, the bottleneck to time-to-science has shifted from compute to data management.



DDN AI400X3. NEXT-LEVEL AI DATA PLATFORM

NVIDIA is the launch customer for AI400X3 with multiple deployments globally in 2025



70% Faster AI Data Throughput

40% DataCenter Savings

PROVEN BY NVIDIA

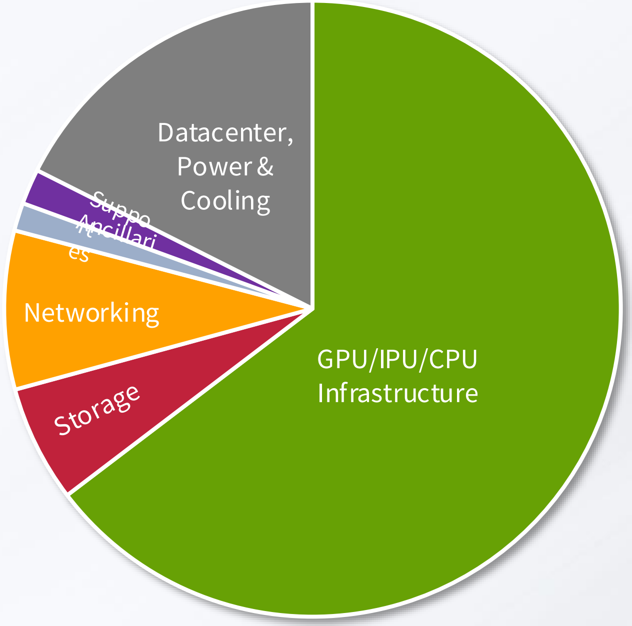
Spend is Large, Risks are High. How can Storage Help?

- A Storage System typically represents 5% of the 3 years Capex & Opex budget of an AI system for Deep Learning/LLM training

- IO Wait and associated elements of the training process can consume up to 43%[1] of runtime

- How can the efficiency architecture and consumption of storage resources impact overall productive output of this System?

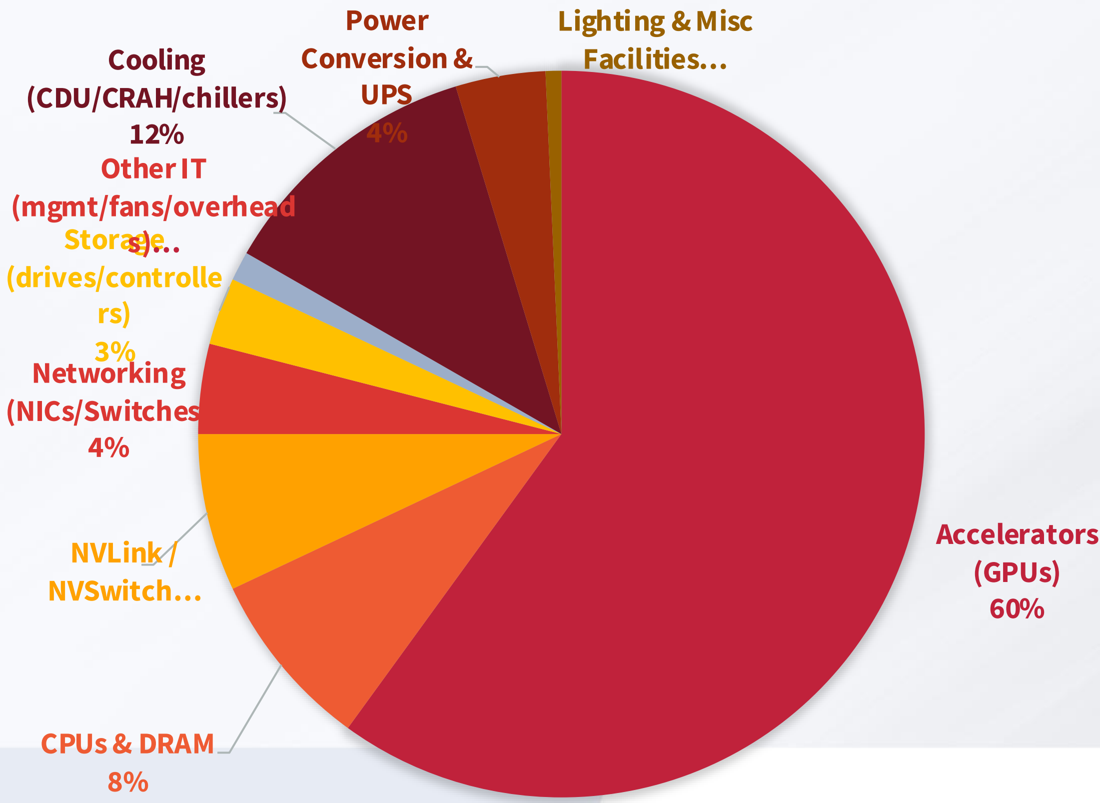
Large AI Supercomputer Capex + Opex



1 (Maeng et al., 2021).

Power Consumption Breakdown (PUE=1.25)

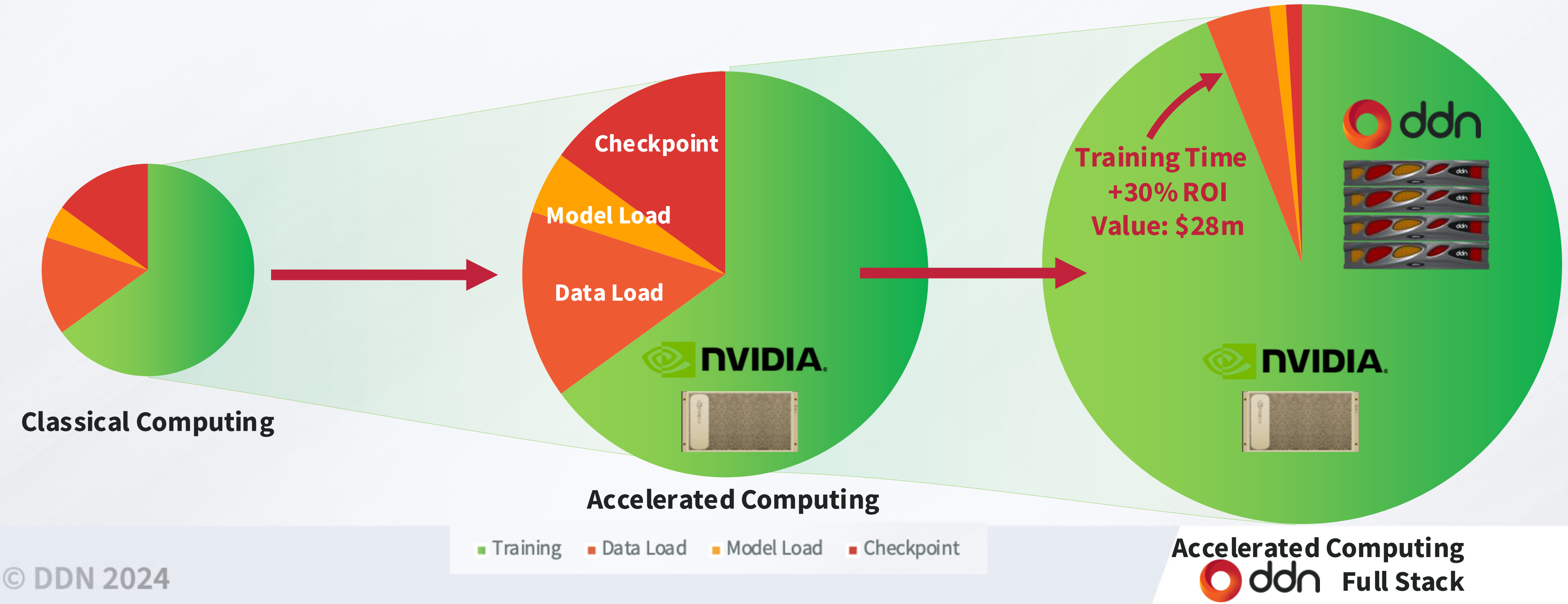
- Storage is only 3%, but can dramatically impact the datacenter efficiency
- Fast DDN storage cuts down the idle/waiting time for GPUs creating more productive output from the datacenter



DDN Enhances Data Center ROI and Profitability by 30%

$1 invested in NVIDIA can translate to $5 in CSP revenue over 4 years

$1 invested in NVIDIA and DDN can translate to $6.5 in data center revenue over 4 years (a 30% increase)



Community Knowledge

IO500: IO patterns vs Storage Systems

- Brought visibility to our community
  - IO as part of the performance equation
- Brought transparency among solutions
  - Mutualization of quantitative data across many sites
  - Fact-Check vendors claims
    - DDN has been supporting the initiative this its inception: if this is not in IO500, it's fishy
- Designed by performance experts
  - On many aspects IO500 is way better than Top500

Production SC23 List

Customize Download

Production | 10 Node Production | Research | 10 Node Research | Full | Historical

Ranking of production system submissions. This is a subset of the Full List of submissions, showing only one highest-scoring result per storage system. Submitters who want a submission that is currently on the Research List to be on the Production List should contact the IO500 Steering Committee.

	INFORMATION							IO500			
#	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE	BW ↑ (GIB/S)	MD (KIOP/S)	REPRO.
1	SC23	Argonne National Laboratory	Aurora	Intel	DAOS	300	62,400	32,165.93	10,066.09	102,785.41	✓
2	ISC23	EuroHPC-CINECA	Leonardo	DDN	EXAScaler	2,000	16,000	648.96	807.12	521.79	✓
3	SC23	LRZ	SuperMUC-NG-Phase2-EC	Lenovo	DAOS	90	6,480	2,508.85	742.90	8,472.60	✓
4	SC23	King Abdullah University of Science and Technology	Shaheen III	HPE	Lustre	2,080	16,640	797.04	709.52	895.35	✓
5	SC23	Memorial Sloan Kettering Cancer Center	IRIS	WekaIO	WekaIO	36	4,248	308.94	104.79	910.80	✓
6	SC23	Japan Agency for Marine-Earth Science and Technology	Earth Simulator 4	DDN	EXAScaler	10	320	101.88	48.19	215.38	✓

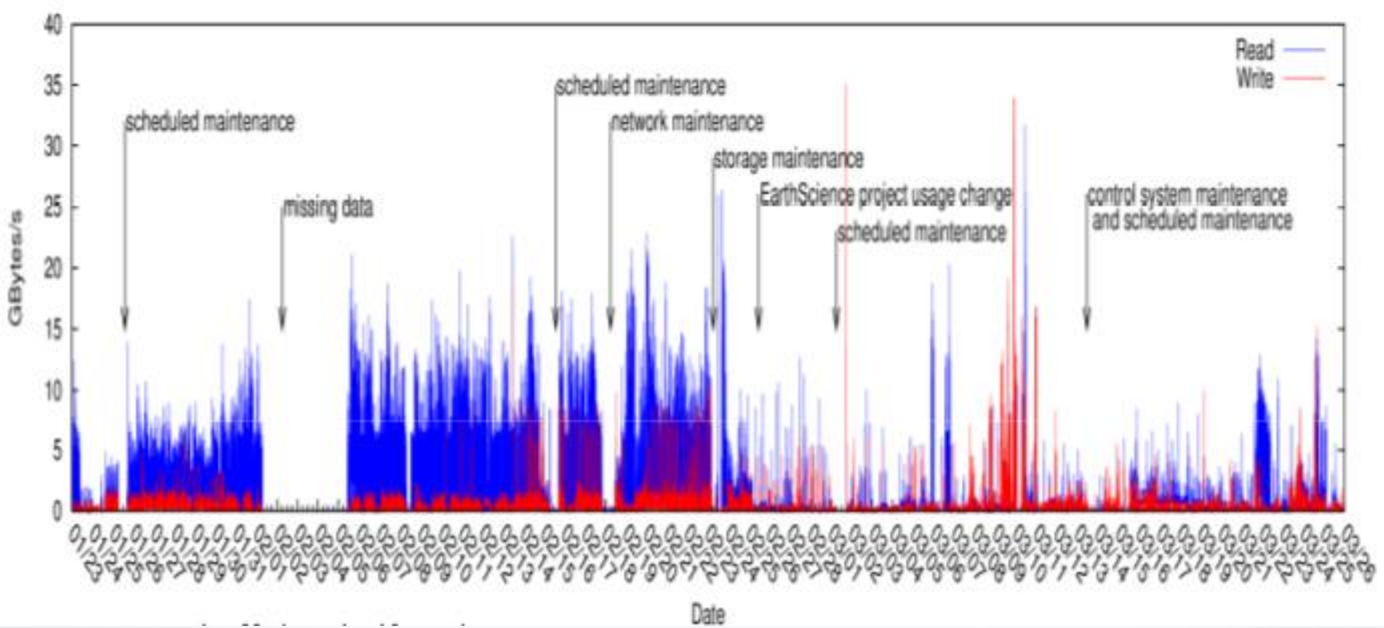
IO500 Design Principles

- IO500 is based on reproducibility of 'meaningful patterns'
- Workload is an evolving landscape
  - Designed for HPC workload
- What used to matter can change
  - High emphasis on Metadata
- What's about I/O libraries?

From System Characterization to workloads Profiling

99% of time IO system stressed less than 33% of its peak bandwidth
70% of time IO system stressed less than 5% its peak bandwidth

Mesures au Argonne National Lab.



“Understanding and Improving Computational Science Storage Access through Continuous Characterization”
PHILIP CARNS et al.
Argonne National Laboratory
2011, Journal Proceedings of 27th IEEE Conference on Mass Storage Systems and Technologies

IO Patterns of key Scientific Applications

HPCIO ANALYSIS

What is a Meaningful Pattern?

- Bottom-Up Approach
  - Addressing the meaning full pattern question
  - End-Users push data
- Per application approach
  - Different settings of the same application generate different IO patterns
    - Parallelism, Check-pointing
- Relies on Standard tool
  - Darshan (Phil. Carnes, Argonne National Lab), de facto standard for I/O tracing
  - Alternative tracing system exists, trace converters are welcomed (OTF2)

HPCIO ANALYSIS

ADMIRE malleable data solutions for HPC

ADAPTIVE MULTI-TIER INTELLIGENT DATA MANAGER FOR EXASCALE

IO-SEA

EuroHPC initiative: IO-Sea + Admire

- Presented at SC'23
  - JGU Group: Prof. André Brinkmann , Nafiseh Moti, Marc-André Vef, Reza Salkhordeh
  - Philippe Deniel CEA , France
  - Jesus Carretero, UCM3, Spain
  - Philip Carns, Argonne National Lab., USA
  - DDN
- Moti, N., Brinkmann, A., Vef, M. A., Deniel, P., Carretero, J., Carns, P., ... & Salkhordeh, R. (2023, November). The I/O Trace Initiative: Building a Collaborative I/O Archive to Advance HPC. In *Proceedings of the SC'23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis* (pp. 1216-1222).
- https://salkhordeh.de/publication/trace-pdsw/trace-pdsw.pdf

Vision

HPC IO Analysis is a community effort aiming at fostering collaboration and improving knowledge of the I/O aspects in HPC applications. Our study aims at creating a shared and open-access database of the I/O performance of applications running on different parallel I/O libraries and in different layers of the I/O stack.
Dealing with large HPC applications on modern infrastructures has become a challenge for most of organizations. This is especially acute for the I/O stack. However, the complexity and heterogeneity of today's scientific and HPC applications make this study challenging. The challenge arises from the different hardware availability and the applications-specific instrumentation efforts; therefore, existing analyses are limited to a few hardware settings and a specific family of applications. An in-depth understanding of I/O characteristics and requirements of different varieties of HPC applications, such as metadata operations and I/O access patterns, is crucial for designing efficient storage and I/O solutions.

The study's comprehensiveness can benefit the whole community of researchers to analyze and spot I/O bottlenecks and further design more efficient system software.

HPC IO Analysis propose to upload I/O traces and will provide in return an analysis. Both the trace and the analysis are made publicly available. Thus, all uploaded data will have to comply with the Creative Common License.

Refer to our **wiki** for an introduction on getting the traces.

Latest Traces

Application	Workload Family	Institute	Cluster/TOP500	IO Library	Checkpointing	MPI Ranks	Actions
DisCoTec, commit e8fad8ae	Burst, Binary MPI-IO	Scientific Computing, Uni Stuttgart	HAWK	MPIIO	False	131072	DETAILS
DisCoTec, commit e8fad8ae	Burst, Binary MPI-IO	Scientific Computing, Uni Stuttgart	HAWK	MPIIO	False	131072	DETAILS

HPCIO ANALYSIS

Detailed information about 65ae57535a23735e2d48cdb0

General Information

Submitter:	Theresa Pollinger
Author List:	Theresa Pollinger, Philipp Offenhäuser
Application Name:	DisCoTec-commit e8fad8ae
Cluster or Top 500:	HAWK
Institute:	Scientific Computing, Uni Stuttgart
DOI:	N/A

DARSHAN FILE

Workload characteristics

Application/Workload Family:	Burst, Binary MPI-IO
Application Link:	https://github.com/SGpp/I
I/O Library:	MPIIO
Checkpointing:	No
Application Size or Link:	https://darus.uni-stuttgart.de/privateurl.xht token=cc65b76d-7d2b-4e43-92f3-c0836587364d (private access token, dataset will be published after paper is accepted)
Additional Information:	striping: 40, num files: 8. With these tests, we investigated the influence of overstriping (the 80 stripes case vs the normal striping at 40), and the number of files written (1 vs. 8 vs. 32)

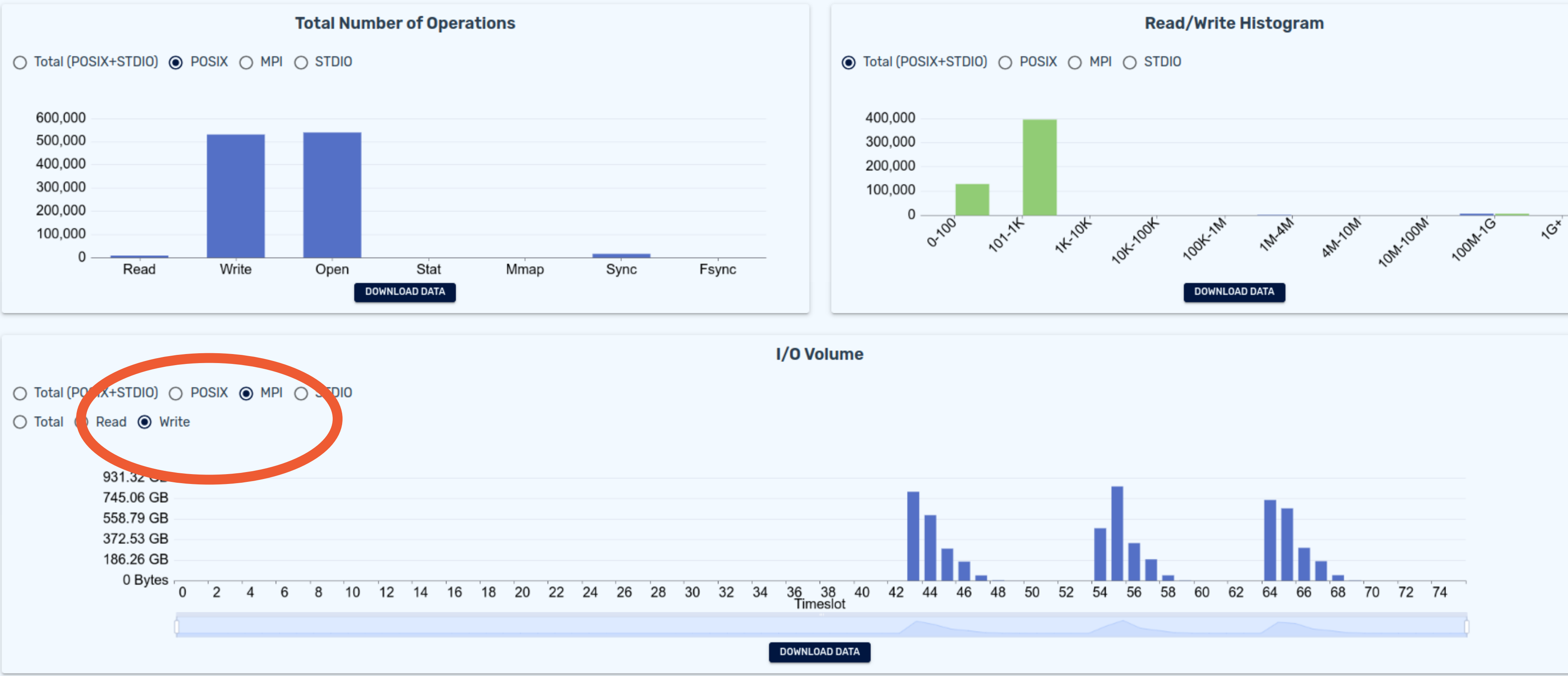
Jobs/environmental information

Number of Ranks:	131072
Runtime:	4 minutes
Slurm Parameters:	cf. DaRUS repository (PBS parameters)

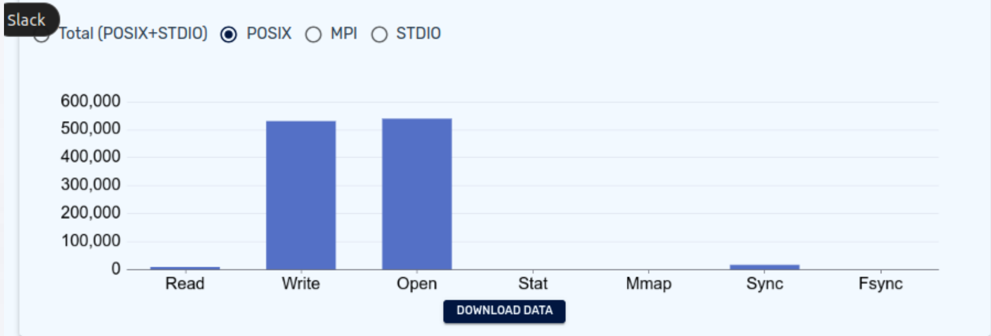
I/O Overview

Reads:	8772
Writes:	561484
Opens:	541588
Stats:	0
Mmaps:	0
Seeks:	16384
Fsync:	0

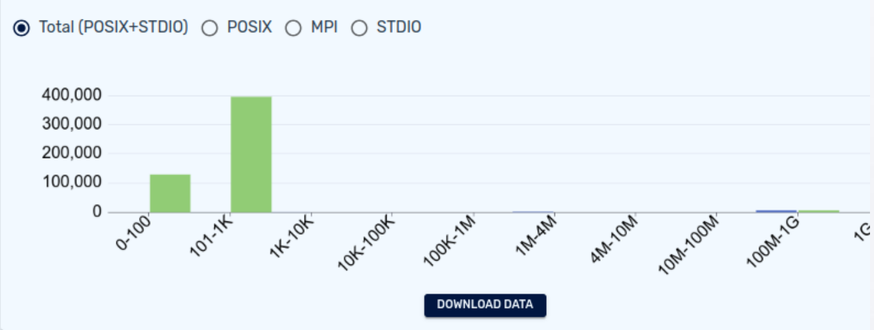
DARSHAN SUMMARY


Total Number of Operations
Total (POSIX+STDIO)
POSIX
MPI
STDIO
600,000
500,000
400,000
300,000
200,000
100,000
0
Read
Write
Open
Stat
Mmap
Sync
Fsync
DOWNLOAD DATA
Read/Write Histogram
Total (POSIX+STDIO)
POSIX
MPI
STDIO
400,000
300,000
200,000
100,000
0
0-100
101-1K
1K-10K
10K-100K
100K-1M
1M-4M
4M-10M
10M-100M
100M-1G
1G+
DOWNLOAD DATA
I/O Volume
Total
POSIX
MPI
Read
Write
745.06 GB
558.79 GB
372.53 GB
186.26 GB
0 Bytes
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74
Timeslot
DOWNLOAD DATA

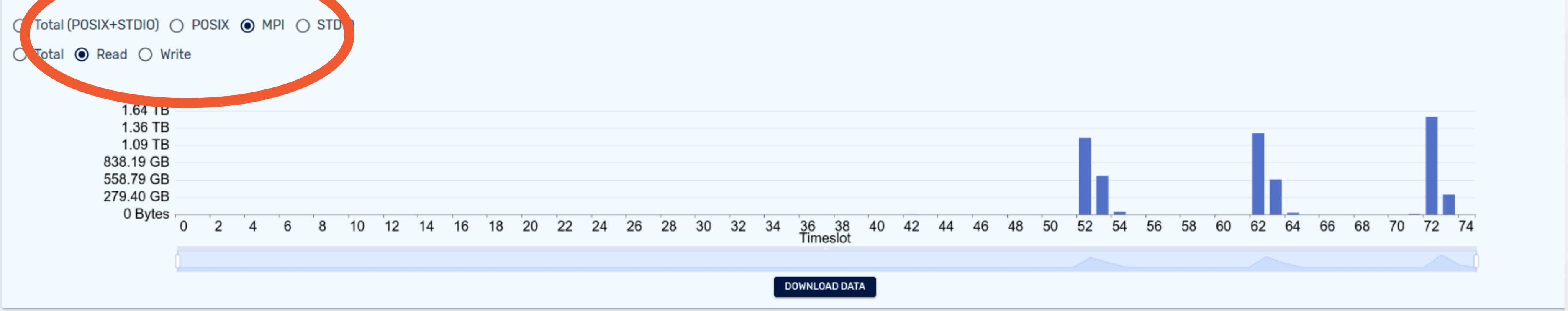
Total Number of Operations



Read/Write Histogram



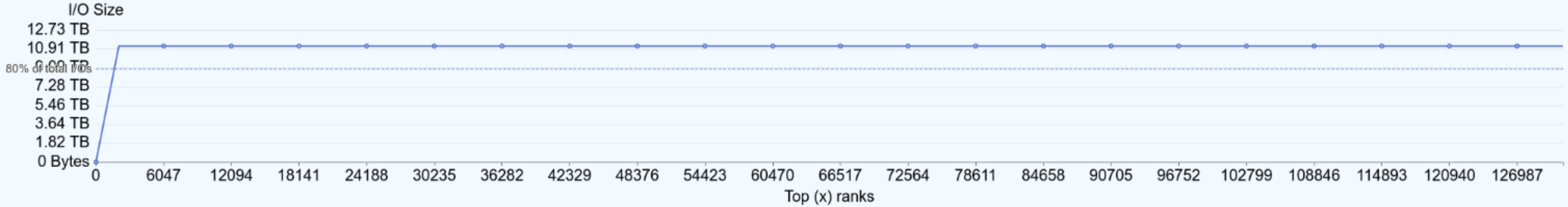
I/O Volume



THE AI DATA COMPANY

Total (POSIX+STDIO) POSIX MPI STDIO

Total Read Write



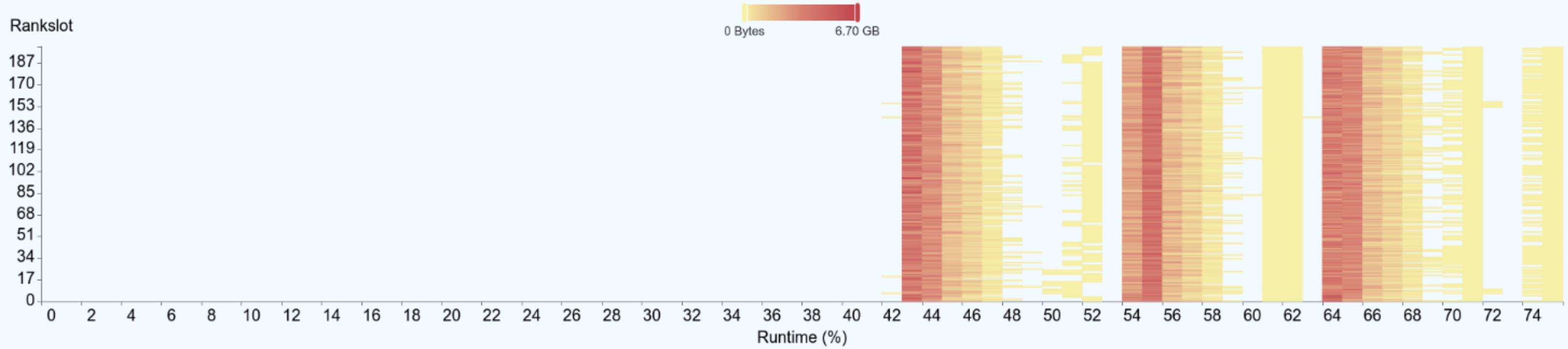
DOWNLOAD DATA

Heatmap

Total (POSIX+STDIO) POSIX MPI STDIO

Normal Logarithmic

Total Read Write



DOWNLOAD DATA

Drishti Analysis

4 critical issue(s) 2 warning(s) 15 recommendation(s)

Metadata Insights

- [P01] Application is write operation intensive (98.37% writes vs. 1.63% reads)
- [P08] Application issues a high number (99.93%) of misaligned file requests
  - Recommendations
    - Consider aligning the requests to the file system block boundaries

Operation Insights

- [P05] Application issues a high number (524288) of small write requests (i.e., < 1MB) which represents 98.84% of all write requests
  - Recommendations
    - Consider buffering write operations into larger more contiguous ones
    - Since the application already uses MPI-IO, consider using collective I/O calls (e.g. MPI_File_write_all() or MPI_File_write_at_all()) to aggregate requests into larger ones
- [P12] Application mostly uses consecutive (5.72%) and sequential (93.14%) read requests
- [P14] Application mostly uses consecutive (0.00%) and sequential (99.95%) write requests
- [P21] Detected write imbalance when accessing 256 individual files
  - Load imbalance of 49.78% detected while accessing "1358049613"
  - Load imbalance of 50.00% detected while accessing "1232763782"
  - Load imbalance of 61.26% detected while accessing "1387466476"
  - Load imbalance of 61.26% detected while accessing "2209925201"
  - Load imbalance of 32.48% detected while accessing "3164901129"
  - Load imbalance of 49.78% detected while accessing "1558704150"
  - Load imbalance of 61.41% detected while accessing "4288096448"
  - Load imbalance of 61.41% detected while accessing "151881144"
  - Load imbalance of 50.00% detected while accessing "2247748081"
  - Load imbalance of 32.48% detected while accessing "3181359701"
  - Load imbalance of 49.78% detected while accessing "1124285698"
  - Load imbalance of 61.26% detected while accessing "4209081358"
  - Load imbalance of 49.78% detected while accessing "4147232947"
  - Load imbalance of 49.78% detected while accessing "1022267489"
  - Load imbalance of 49.78% detected while accessing "1591958127"
  - Load imbalance of 49.78% detected while accessing "122752989"
  - Load imbalance of 32.48% detected while accessing "1308684948"
  - Load imbalance of 50.00% detected while accessing "2130329966"

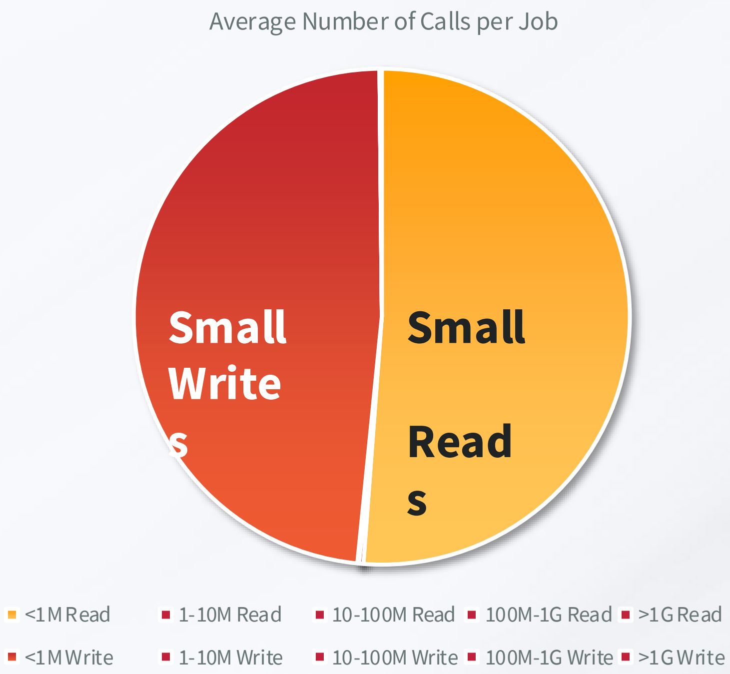
HPC IO Analysis Needs You!!

- Algorithms are mostly optimized for compute
  - Better exposition of I/O patterns will eventually lead to better applications
- End-Users do no to fully grasp the impact of their parameterization
  - Data volume is not the only, and not the most important, explanatory variable
- Sys. Admin will have more detailed information
  - Configuration of the system
- Vendor use public data to design their storage solutions

Machine Learning is Write and Read Intensive

- Analysis of over 23,000 Machine Learning Jobs

- “Most ML jobs are perceived to be read-intensive with a lot of small reads while a few ML jobs also perform small writes.”
- “Our study showed that ML workloads generate a **large number of small file reads and writes**…”



Characterizing Machine Learning I/O Workloads on Leadership Scale HPC Systems
https://arnabkrpaul.github.io/publications/mascots21.pdf

DDN Accelerates the Thousands of Checkpoints Needed in AI

Prediction Accuracy – Improve accuracy by lowering learning rate from a checkpoint

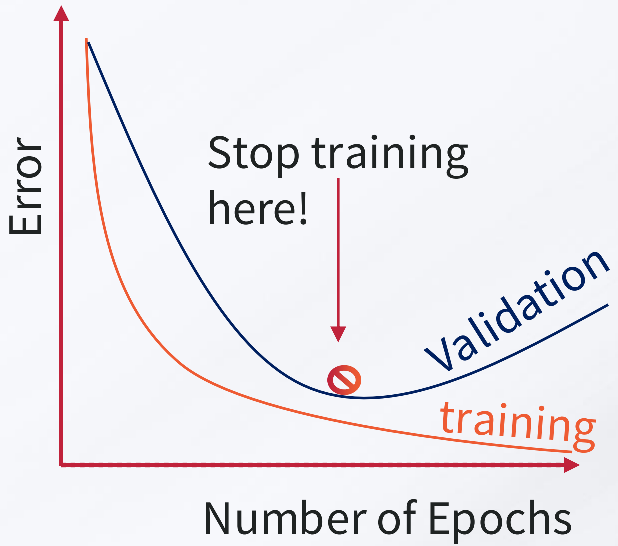
Multi-System Training - continue training model across different nodes or clusters/cloud

Transfer Learning – if goals change, start afresh from a checkpoint

Better Fine Tuning - pick out less trained states to restart new experiments

Early Stopping - For large models, without sufficient regularization, the error on the evaluation dataset can start to increase.

- → need to go back and export the model that had the best validation error.



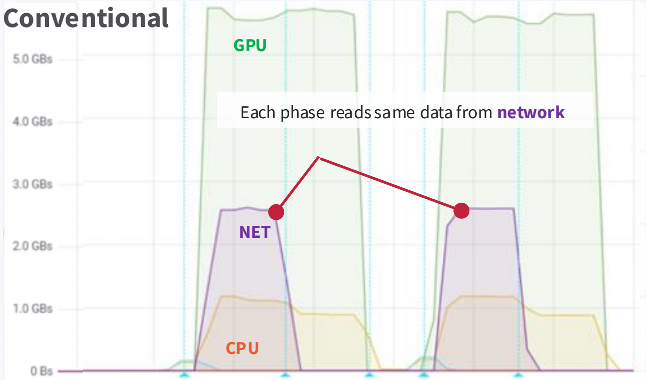
Checkpoints is intrinsic to Deep Learning Training

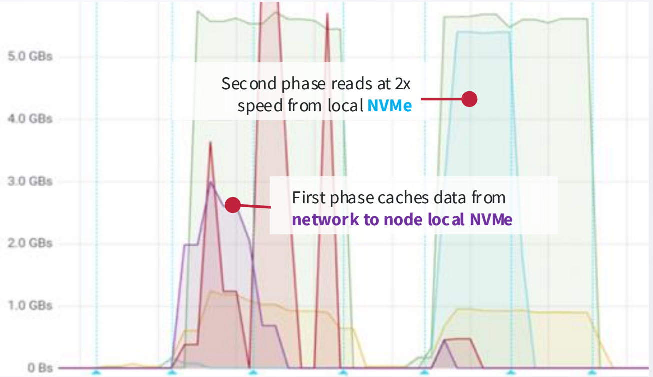
- **Non-linear convergence** – local minimum exist where the next epoch degrades accuracy but on the long run accuracy can still be improved
- **Over-fitting to detect Sweet-Spot –** some overfitting is mandatory to ensure that the global minimal has been reached
- **Rolling-back to the Global minimum** – once the detection of the global minimum has been assessed, rolling back to correct model state requires parsing the checkpoint history

Optimizing Multi-Epoch Training

- Without DDN Hot Nodes technology, Multi-Epoch Training consumes storage and network bandwidth with every GPU systems repeatedly pulling data.
- With DDN Hot Nodes, we automatically cache data sets on internal NVMe devices, freeing the network and storage from load and accelerating the whole training process

Conventional





HOT NODES

MLPerf Storage for AI Workloads

MLPerf Storage targets AI

- Address the loophole of AI
- AI is complicated
  - Field is constant evolution
    - From read driven to write driven
  - Out of core aspect
- Training and inference subtleties
- Training is extremely expensive

MLPerf Storage v2.0 - Workloads for training

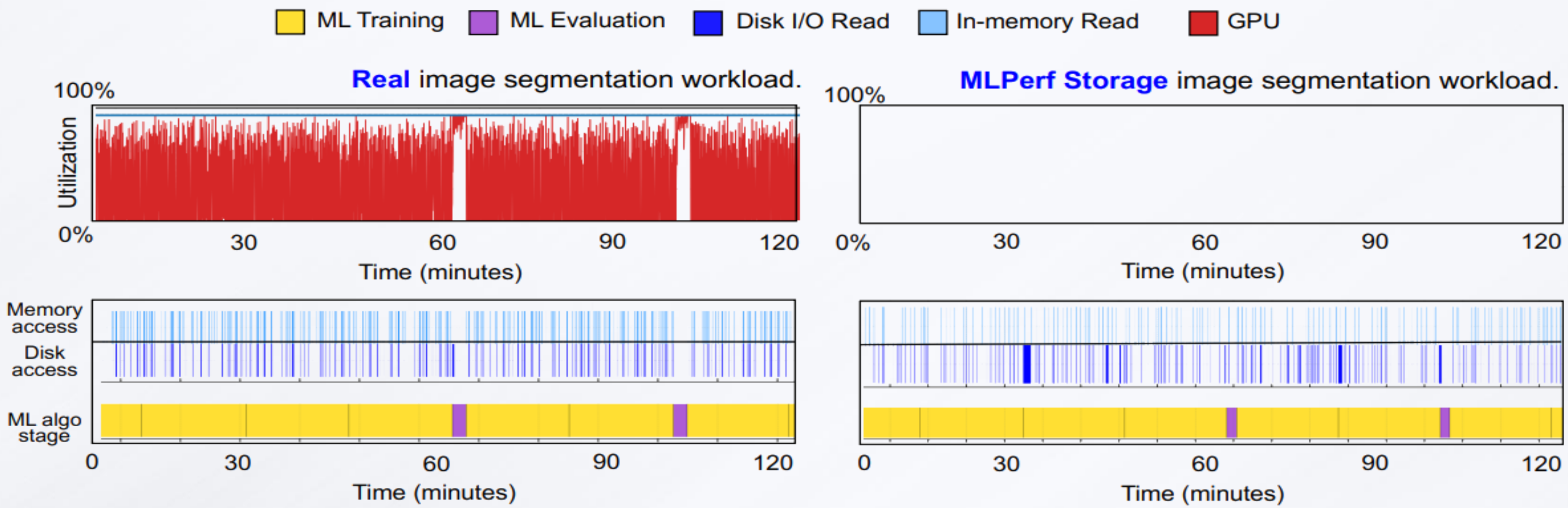
Measures how fast storage systems can supply training data when a model is being trained.

MLPerf Storage ***emulates*** GPU performance, specifically assessing the I/O components of AI training protocols.

A sleep() function is called to simulate the compute part. **No real GPU is needed.**

Metric	V0.5	V1.0	V2.0
# submissions	24	150	231
# submitters	5	17	~30

MLPerf Storage v2.0 - Workloads for training



From Characterizing Machine Learning I/O with MLPerf Storage
***Oana Balmau** - CHEOPS @ EuroSys, May 8th, 2023*

Area	Problem	Model	Data Loader	Dataset seed	Minimum AU%
Vision	Image segmentation (medical)	3D U-Net	PyTorch	KiTS 19 (140MB/sample)	90%
Vision	Image classification	ResNet-50	TensorFlow	ImageNet (150KB/sample)	90%
Scientific	Cosmology	parameter prediction	TensorFlow	CosmoFlow N-body simulation (2MB/sample)	70%

MLPerf™ Storage v2.0 – Comparison w/ X2T (throughput)

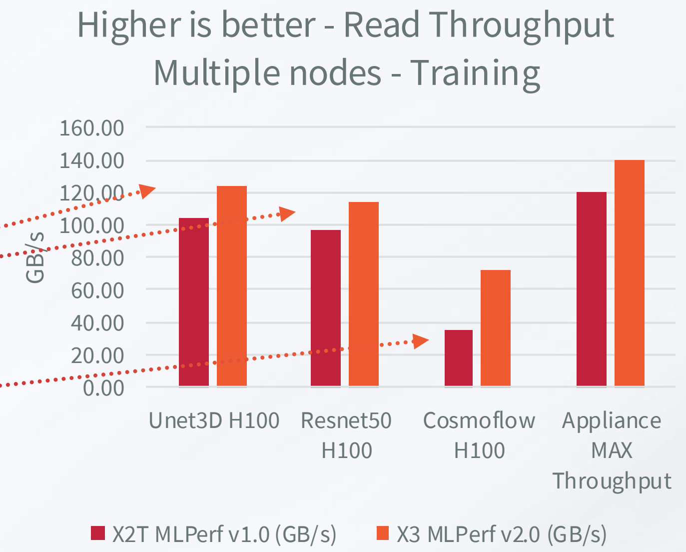
- **Multi. Host - CLOSED**

Training H100	Unet3D (GB/s)	Cosmoflow (GB/s)	Resnet50 (GB/s)
X2T (v1.0)	103	96	35
X3 (v2.0)	123	114	71

Performance improvement due to better appliance throughput:

Performance improvement due to better tunings:

- **Leverage logs from v1.0 from YanRong/other submitters for MLPerf Storage tunings**
- **Zero lustre preload (important improvement for epoch 0)**
- **RPC_In_flight=1/client (latency limited workload)**



Energy and Performance

Performance and Power: looking from the PDUs

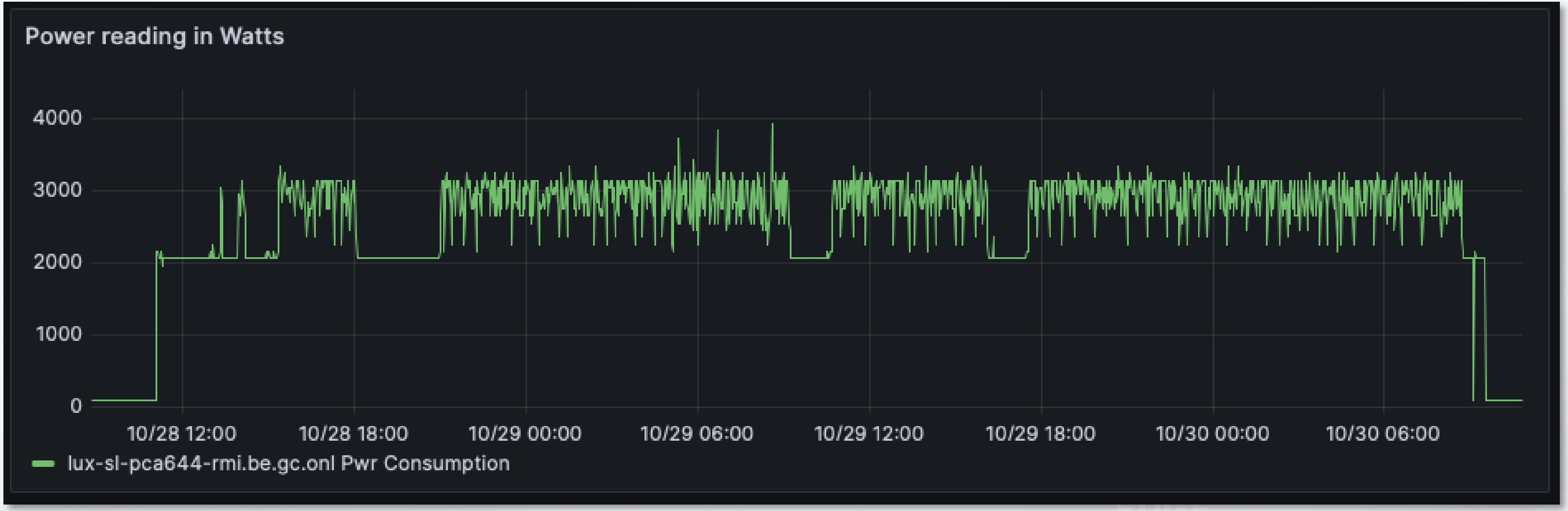
Performance and Power: I/O Pattern impact

THE AI DATA COMPANY

I/O pattern	idle Power (Watt)	Peak Power (Watt)	BW (GB/s)	IOPS
Write sequential 4M	430	**564**	**40**	10K
Write sequential 4KB	430	565	35	9260K
Write Random Write 4BK	430	470	1.2	320K
Write Single Shared File Random 4KB	430	470	0.03	7500
Read sequential 4M	430	**510**	**48**	12K
Read sequential 4k	430	510	25	6.5M
Read Random 4k	430	455	1.7	430K

Performance and Power: the Whole Picture

THE AI DATA COMPANY



8-A100 GPU node. Courtesy of G-CORE lab

NVIDIA Reference architecture: 64 GPU per Storage appliance

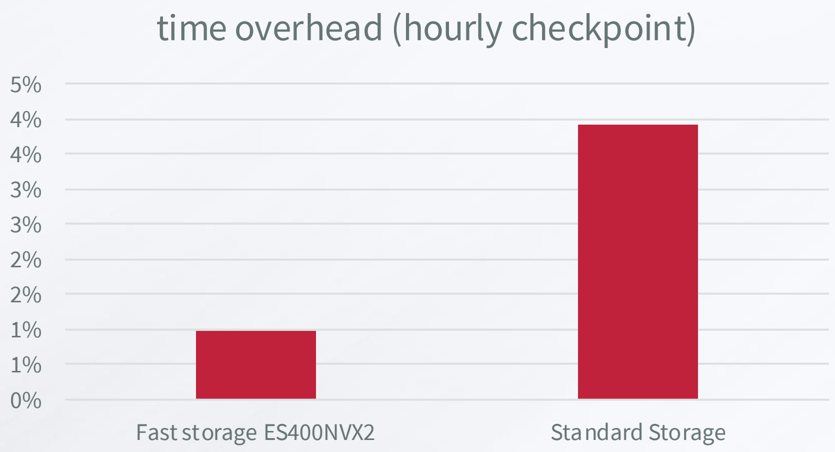
- GPU 16-24KW
- Storage 0.8 to 1.3 KW

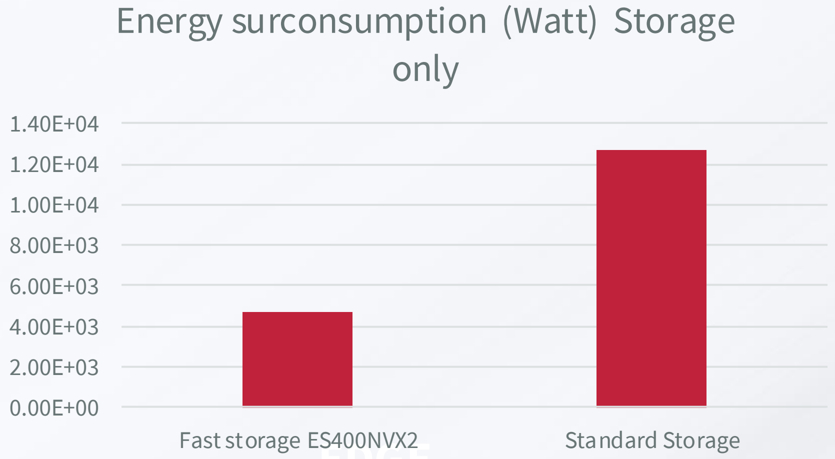
Fast storage saved 8KW on storage, and save 1.7MW of GPU

Performance and Power: the Whole Picture

Fast storage: 40 GB/s write

Standard storage: 10 GB/s write





Checkpointing (e.g., Adam, most common for large models):

176 billion parameters * 2 (moments) * 4 bytes/moment (FP32) = 1408 GB (1.408 TB)

Data and Metadata

Data and Metadata

Metadata : data describing data

- data attributed (and extended attribute)

→ size, user access rights, date of modification (ls –l)

!= pointer (does not allow to locate data on the storage system)

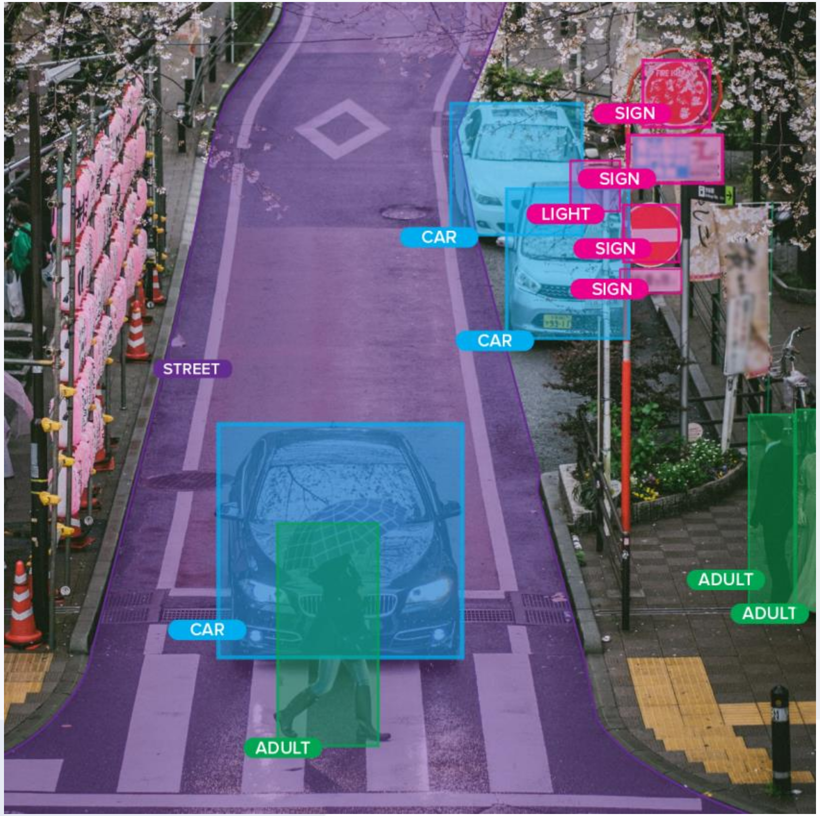
Metadata are at the core of the scalability challenge

→ metadata are accessed frequently and massively, e.g. *ls* –l in a directory

with many file. No data access but many metadata accesses

Blurring Borders: Metadata & Data

- AI Data tend to be metadata heavy
  - Every frame of an autonomous car is annotated by 100s of metadata
- Metadata allow to structure the Data-lake
  - Prevent Data-lake to turn in Data-Swamp
- Query-able Metadata: Data-LakeHouse
  - Data Lake + Data Warehouse

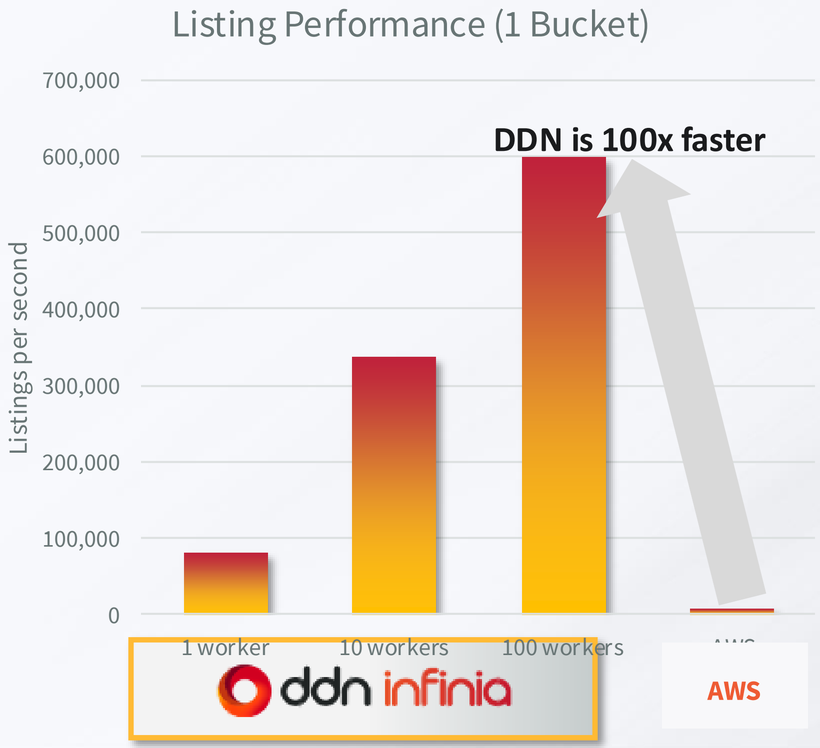


Fast Object Listing + SQL = Dataset selection

```
-- Satellite images by time range
Sql> SELECT * FROM Copernicus WHERE date >= '2025-01-01' AND
date < '2025-02-02';

-- Satellite images for a given area and time range
Sql> SELECT * FROM Copernicus WHERE date >= '2025-01-01' AND
date < '2025-02-02' AND latitude BETWEEN min_latitude AND
max_latitude AND longitude BETWEEN min_longitude AND
max_longitude;

-- Satellite images for a given area and time range and
a specific feature
Sql> SELECT * FROM Copernicus WHERE date >= '2025-01-01' AND
date < '2025-02-02' AND latitude BETWEEN min_latitude AND
max_latitude AND longitude BETWEEN min_longitude AND
max_longitude AND water_body_presence > 0.85;
```



```
ubuntu@ip-172-31-31-36:~$ ./go-s3-tests list million --max-keys 5000 --perf --expected-objects 1000000
```

ddn infinia

ubuntu@ip-172-31-31-36: ~

DDN Infinia

```
AWS S3 Express One Zone

ubuntu@ip-172-31-31-36:~$ ./go-s3-tests list million--use1-az4--x-s3 --use-vhost --max-keys 5000 --perf --expected-objects 1000000
```

AWS S3 Express

Latency

Technologies tend to stack

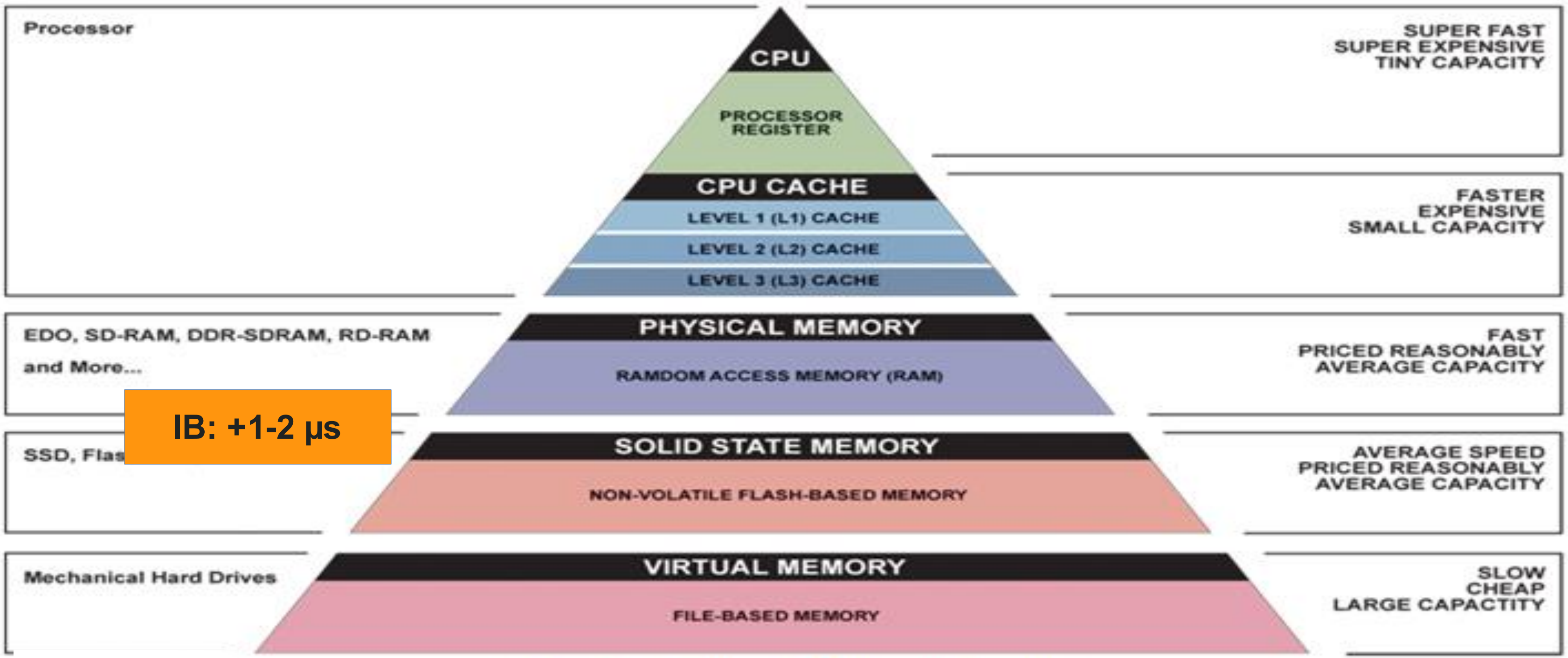
1-10 ns

50-100 ns

100 µs

5 ms

50 sec



IB: +1-2 µs

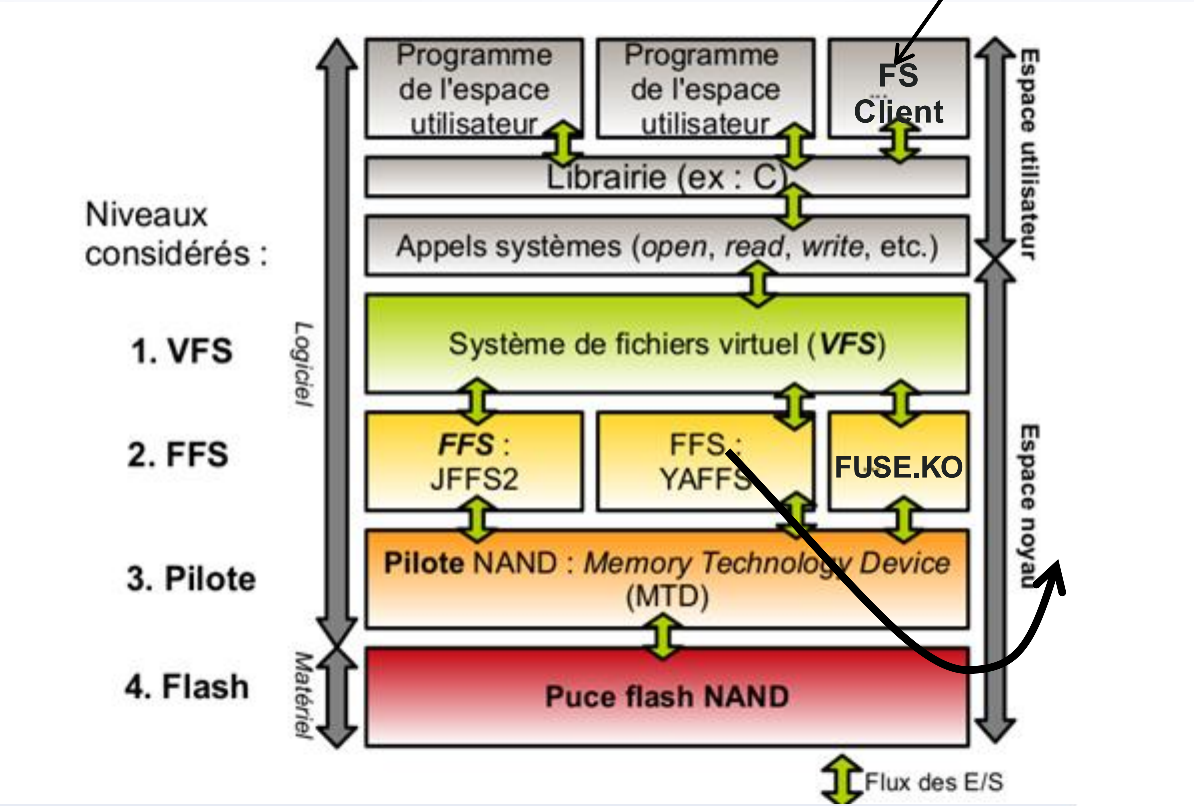
▲ Simplified Computer Memory Hierarchy
Illustration: Ryan J. Leng

Key Performance metrics: devices

Metrics	Hard Drive (HDD)	Flash (NMVe) PCI gen4
Bandwidth	0.2 GB/s	8 GB/s
Latency	4 ms	0.02ms
Capacity	22 TB	60 TB (QLC)
Price	$14.3 / TB	$50 / TB

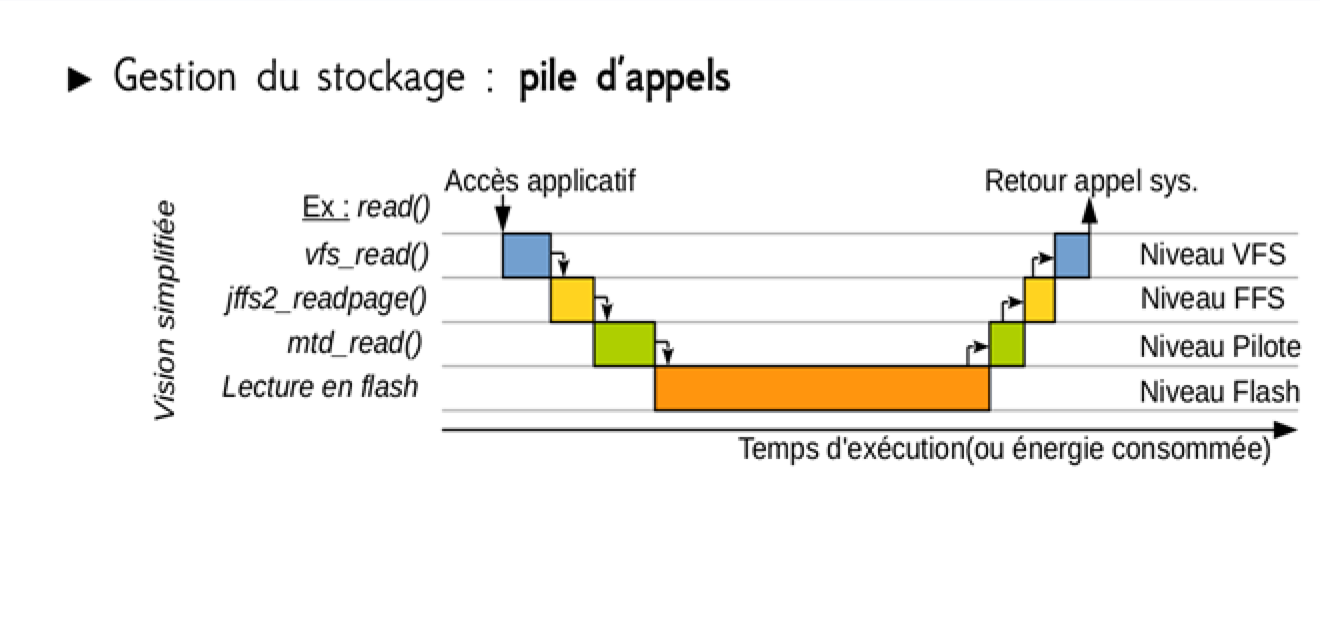
Call Stack

Management of memory cache and network FS communication protocol



Source: P. Olivier et J. Boukhobza

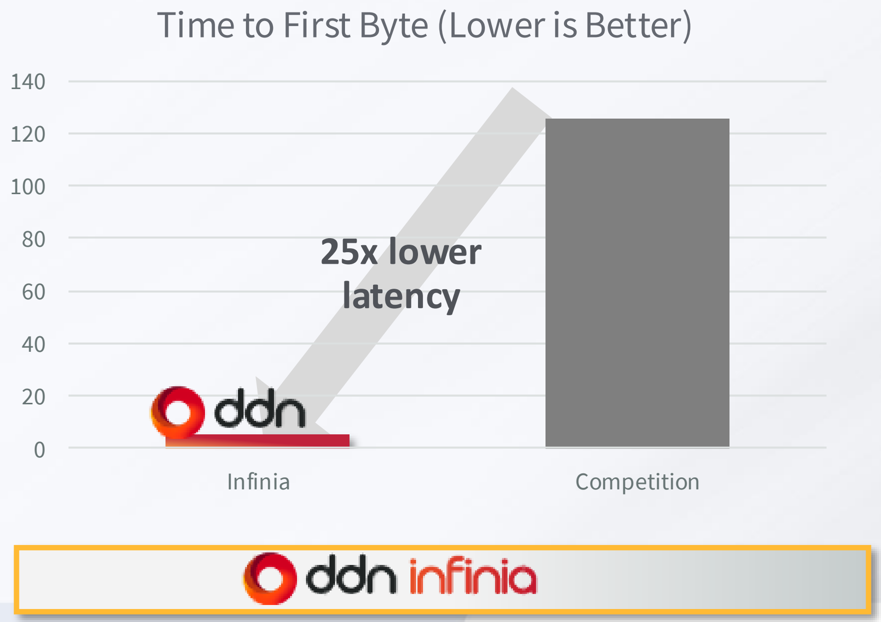
Overhead associated to Call Stack



Source: P. Olivier et J. Boukhobza

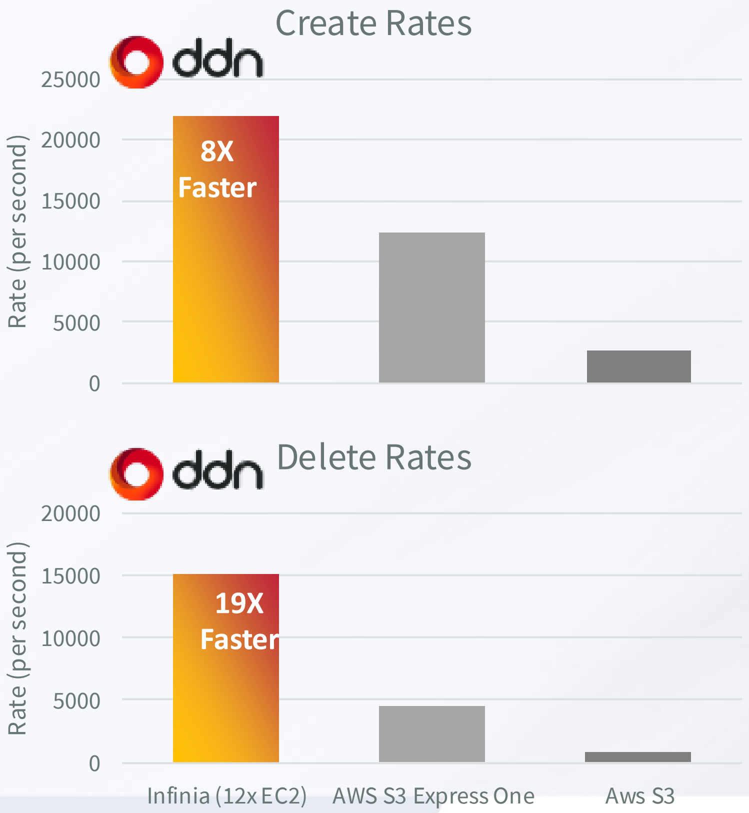
25X Faster Response Times for Data Access

- Rapid data retrieval, enables real-time interactions critical for RAG apps
- **Enhanced User Experience**: empowers seamless, high-performance workflows, improving end-user satisfaction and business outcomes.



Small Object is latency driven

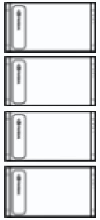
- DDN Infinia running on Public Cloud outpaces Native Object Storage for the most common operations



From Applications to Workflows

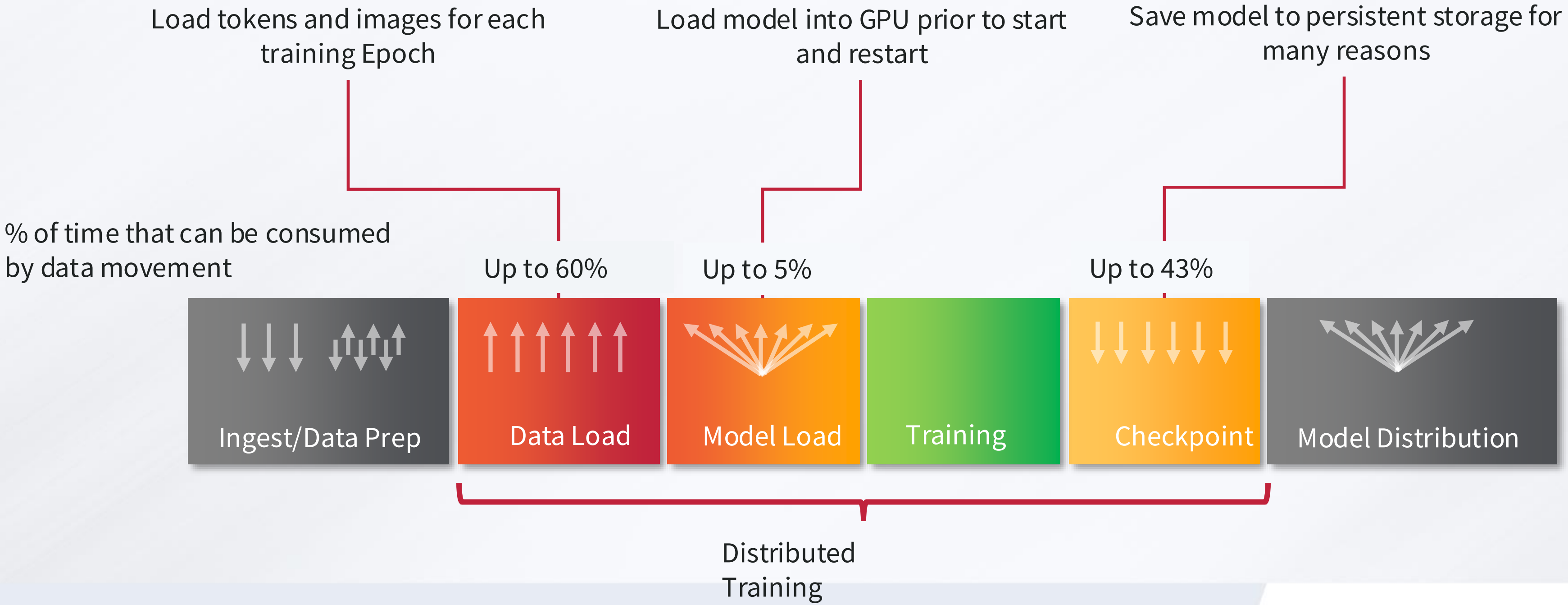
What does the End-to-end data journey involve?

Example: Life Sciences – Drug Discovery Application

CAPTURE	PREPARATION	PROCESS	ANALYZE	ARCHIVE
				
Millions of digital images captured	Cleansing and labeling of data	Model training Computation and analysis	Examine & Explore results for action	Store data for long-term compliance

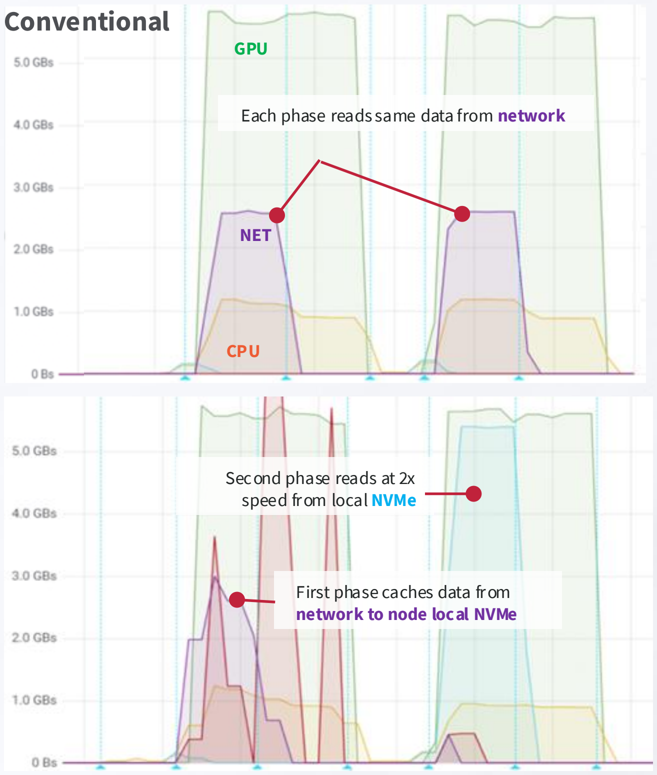
End-to-end data journey

AI data life cycle goes beyond training



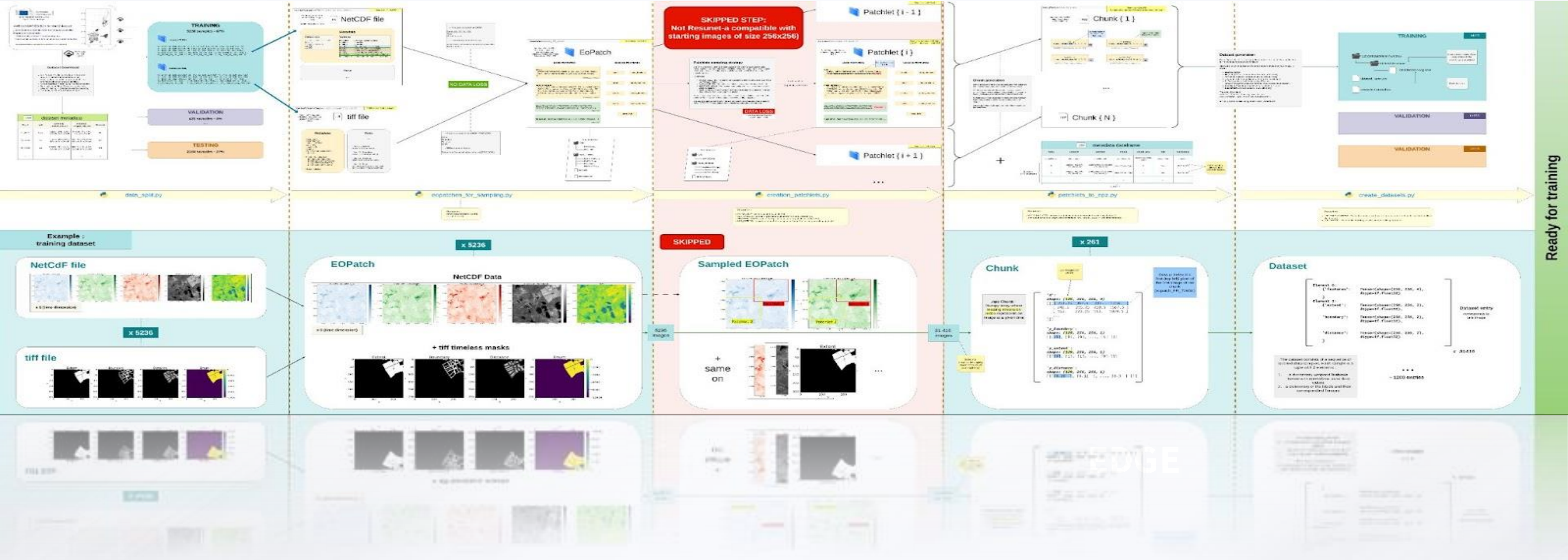
Optimizing Multi-Epoch Training

- Without DDN Hot Nodes technology, Multi-Epoch Training consumes storage and network bandwidth with every GPU systems repeatedly pulling data.
- With DDN Hot Nodes, we automatically cache data sets on internal NVMe devices, freeing the network and storage from load and accelerating the whole training process



HOT NODES

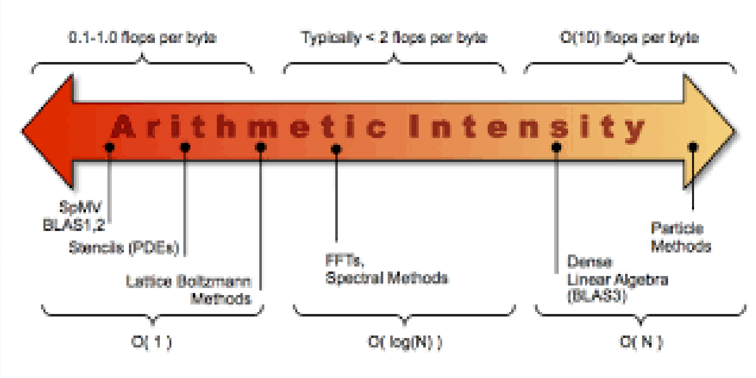
Roof line for Workflows 1/2



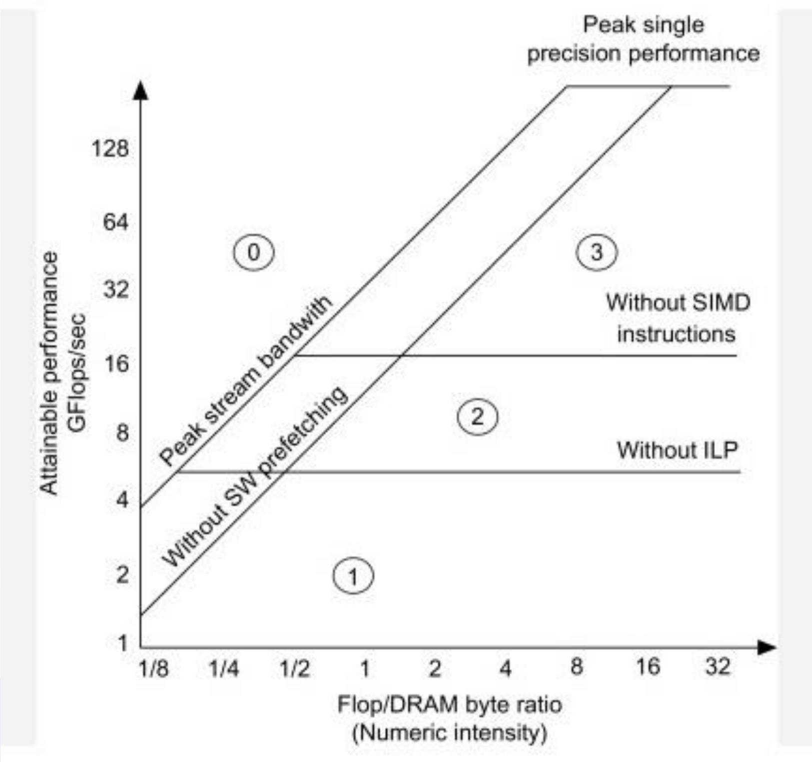
Each Stage has its own arithmetic intensity

Roofline model for Coarse Grain Characterization

Intrinsic Application Characteristics



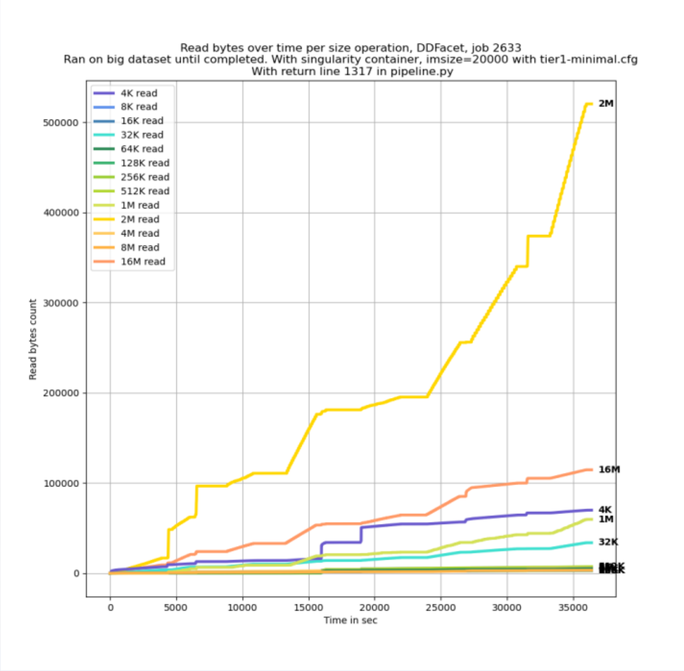
Application + Platform Characteristics



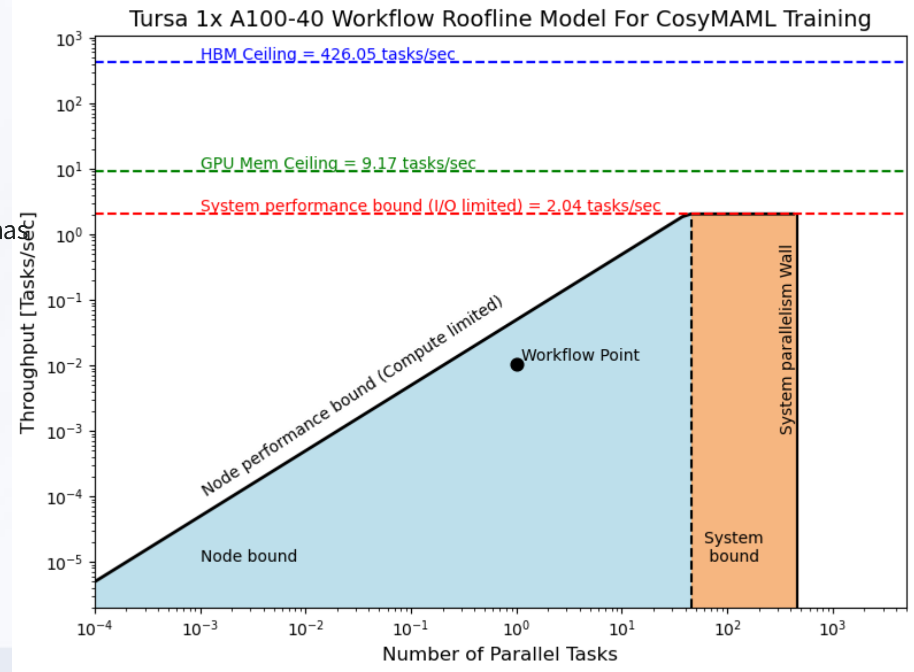
Roofline for Workflows 2/2

Ding, Nan, Brian Austin, Yang Liu, Neil Mehta, Steven Farrell, Johannes P. Blaschke, Leonid Oliker, Hai Ah Nam, Nicholas J. Wright and Samuel Williams. "A Workflow Roofline Model for End-to-End Workflow Performance Analysis." SC24, https://crd.lbl.gov/assets/Uploads/Workflow_roofline-6.pdf

$$TPS \leq \min \begin{cases} \dfrac{\text{Number of parallel tasks}}{\text{Bytes\_sys / Peak sys GB/s}} \\ \dfrac{\text{Number of parallel tasks}}{\text{Bytes\_node / Peak node GB/s}} \\ \dfrac{\text{Number of parallel tasks}}{\text{Flops\_node / Peak node TFLOPS}} \\ \textit{other ceilings not subjected to the above} \end{cases}$$



has



Lustre job stats

Stage roofline

Upcoming I/O Patterns: Inference

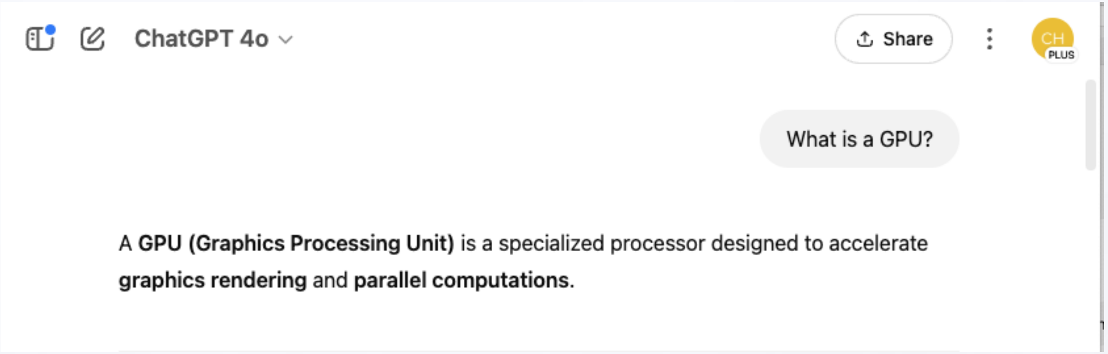
A day in the life of a Prompt

THE AI DATA COMPANY

Query 1

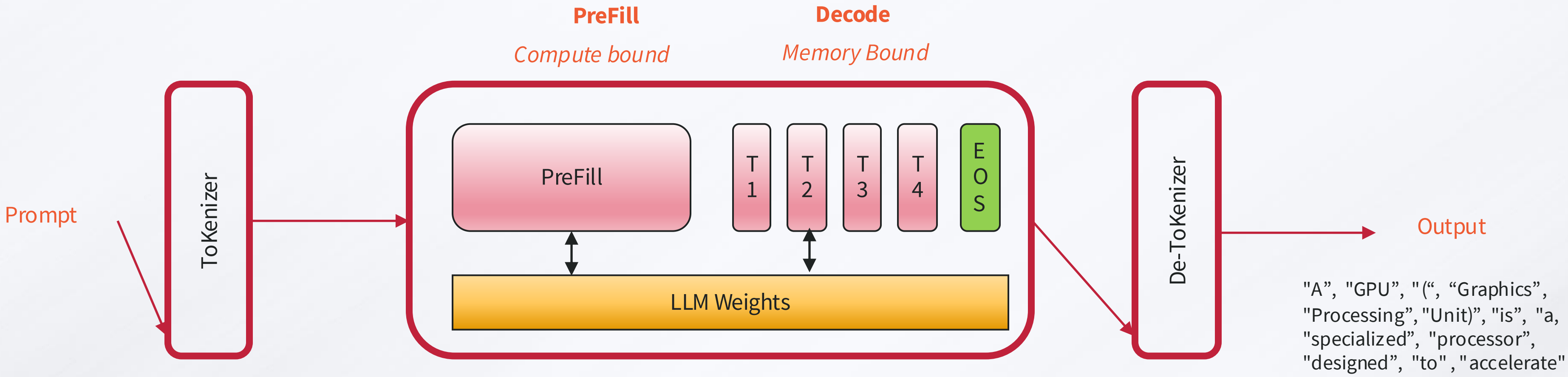
Response 1

Return



Transformer Workflow – Tokens >> Prefill >> Decode >> Output

Autoregressive process that generates each token based on previous Tokens.



System Prompt	13718	480	14567	827	14567
User Prompt	What	Is	A	GPU	?
	4827	382	261	47969	3901

Process the entire input sequence to initialize the **KV cache** for efficient autoregressive decoding.

What is a KV cache?

A memory mechanism inside LLMs that stores Key (K) and Value (V) representations of previous tokens.

T1	T2	T3	T4	T5

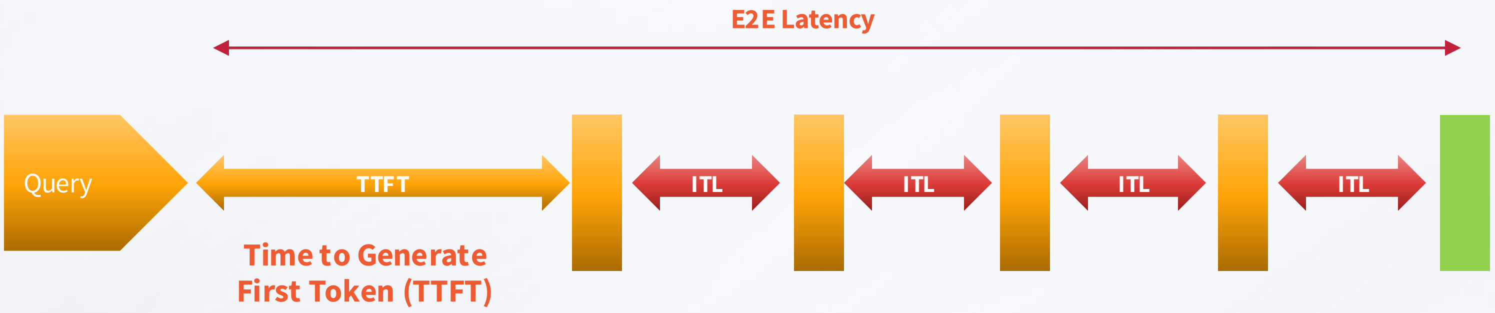
Why It Matters— Helps with Token Generation. Captures the Computation – semantics, position, attributes

- **Autoregressive decoding**, every new token depends on previous ones.
- **Prevents precomputation** of all past tokens, making token generation much faster and more efficient.

Key Challenge:

•Memory-Hungry - Grows sequence length and batch size.
•Primary Consumer of GPU memory

SLAs under consideration



Input sequence length of tokens (ISL) are fed into a model.

- Time it takes for the model to generate the first token after receiving a request. **Prefill**.
- Compute-bound

- **Inter-token Latency (ITL)** is an average time between output tokens.
- Critical when full response has to be processed further: guardrails, tool calling, Agent Calling
- Memory-bound

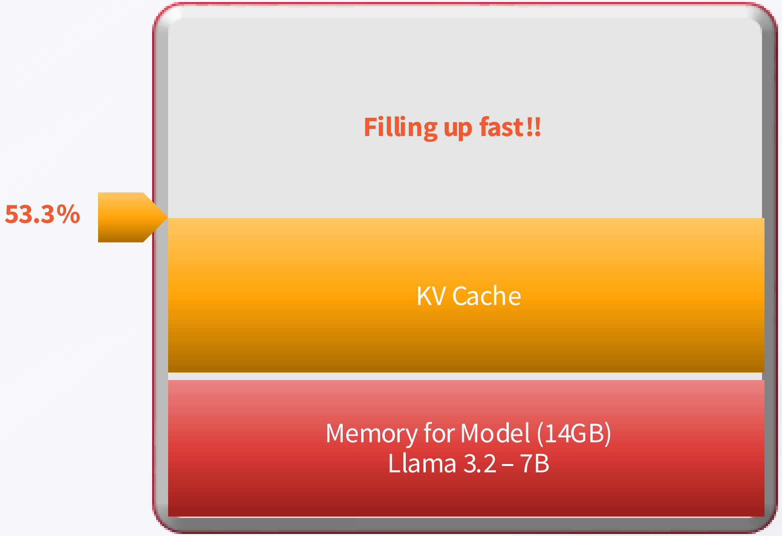
LLM generates *output sequence length (OSL)* of tokens one at a time.

GPU Memory gets filled up past

Memory Profile - 4 Users, Small Query, small model, Quantized

Batch Size	Model Memory (GB)	KV Cache Memory (GB)	Total Memory (GB)	KV Cache % of Total
1	14	4	18	22.2%
2	14	8	22	36.4%
3	14	12	26	46.2%
4	14	16	30	53.3%

NVIDIA H100 – 80 GB of Memory

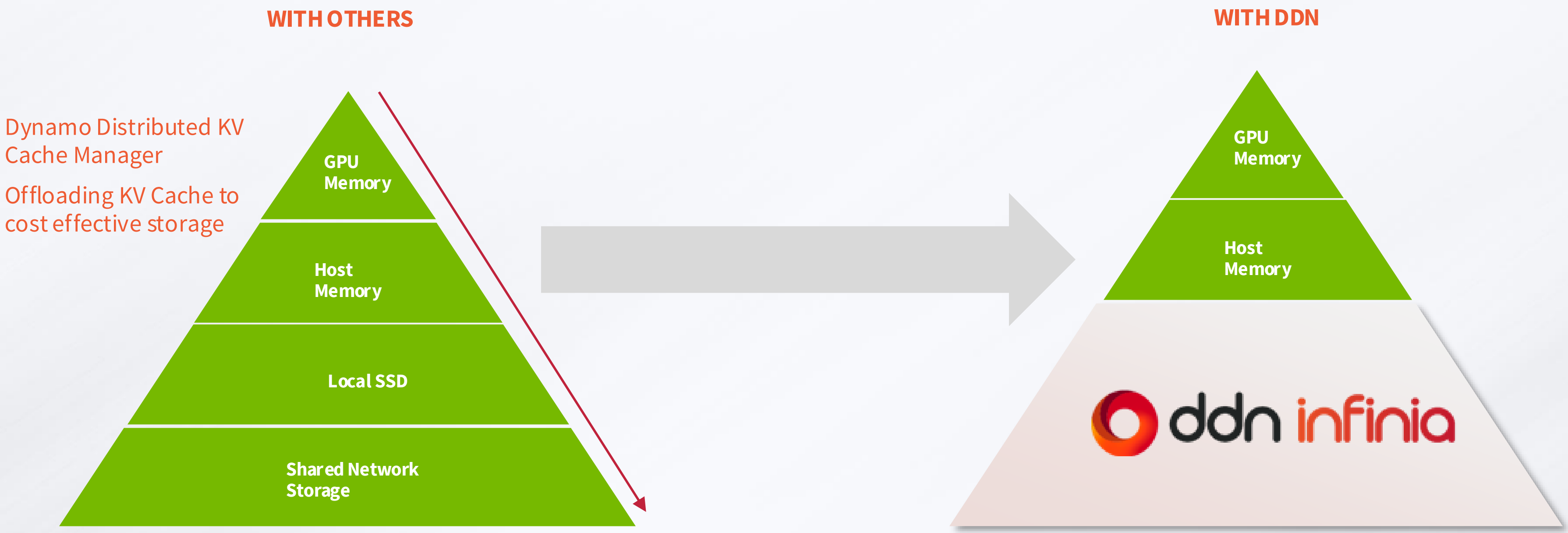


Batch Size = 4, Seq Len - 2048

Infinia's Low Latency KV Architecture Make it Ideal for Managing Distributed KV Securely

Feature	How Infinia Helps
Disaggregated Cache	Infinia stores and serves the cache across GPUs
Parallel Access	Multiple GPUs can read/write to the cache simultaneously
Low Latency	NVMe-oF ensures fast access to cached data
Scalable	Infinia scales horizontally and vertically
Fault Tolerance	KV cache state persists even if GPUs fail
Efficient Caching	Tiered storage keeps "hot" data close to GPUs

DDN Infinia KVCache Acceleration for Inference



2. Models are bigger, wider context window & Attention Heads – More GPU memory and KV Cache Size

Model Name	Organization	Year Introduced	Model Size	# Parameters (B)	Context Window (tokens)	# Attention Heads
GPT-4 Turbo	OpenAI	Late 2023	Large	~1,000*	128,000	96* (est.)
Gemini 1.5 Pro	Google	2024	Large	Undisclosed*	1,000,000+	Undisclosed
Claude 3 Opus	Anthropic	2024	Large	Undisclosed*	200,000	Undisclosed
Llama 3 70B	Meta	2024	70B	70	8,192	64
Mistral Large	Mistral AI	2024	Large	Undisclosed*	32,000	Undisclosed
Yi-34B	01.AI	Late 2023	34B	34	32,000	52
Qwen 2 72B	Alibaba	2024	72B	72	128,000	72
DBRX	Databricks	2024	132B	132	32,000	96