REACHING JUPITER: MISSION DETAILS AND LAUNCH PREPARATIONS

JLESC24 Kobe

2024-04-17 I ANDREAS HERTEN | JÜLICH SUPERCOMPUTING CENTRE

Member of the Helmholtz Association

EuroHPC Joint Undertaking

Bundesministerium für Bildung und Forschung

Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen

JÜLICH Forschungszentrum

Shaping Change

Mission Briefing 2024-Apr-17

Introduction for new travelers

MISSION HISTORY

THE EUROPEAN ROAD TO EXASCALE

- 2011 • First Exascale-focused EU Projects funded by FP7: DEEP, Montblanc, CRESTA
- 2015 • Discussion about direct funding of HPC-Infrastructure in the EU
- 2018 • EuroHPC JU is founded
  GCS/BMBF presents their “Smart Scaling Strategy”
- 2021 • Call for Expression of Interest for Hosting Entities for 1st Exascale system in Europe
- 2022 • Pre-Exascale systems Lumi and Leonardo are installed
- 2023 • Pre-Exascale system MareNostrum 5 is installed
- 2024 • Exascale: JUPITER

JSC

BECOMING A HOSTING ENTITY/SITE

Call for Expression of Interest to host a high-end supercomputer by EuroHPC JU
17.12.2021

Hearings
16.05.2022

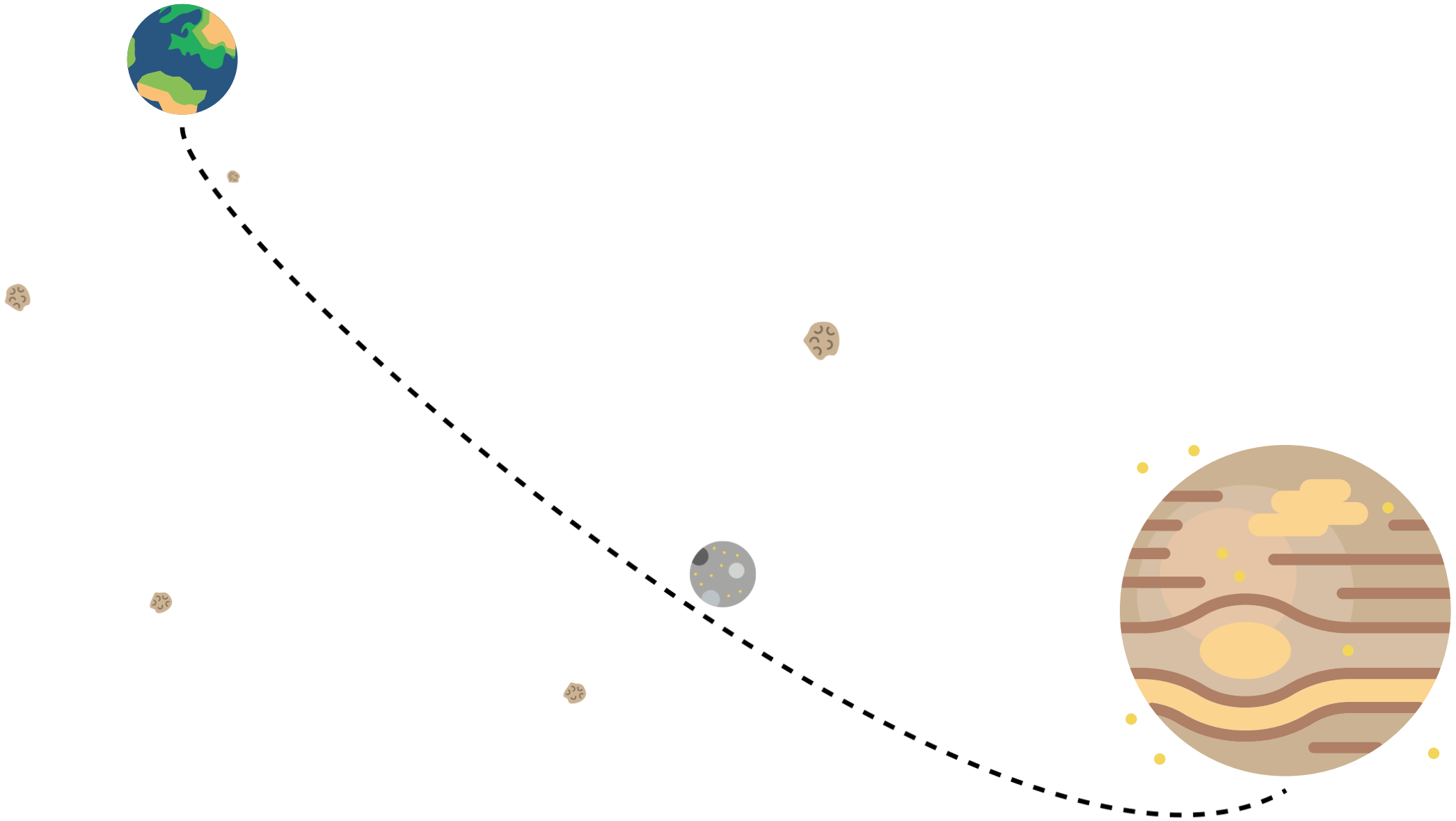
December → January → February → March → April → May → June

Deadline for Application
14.02.2022

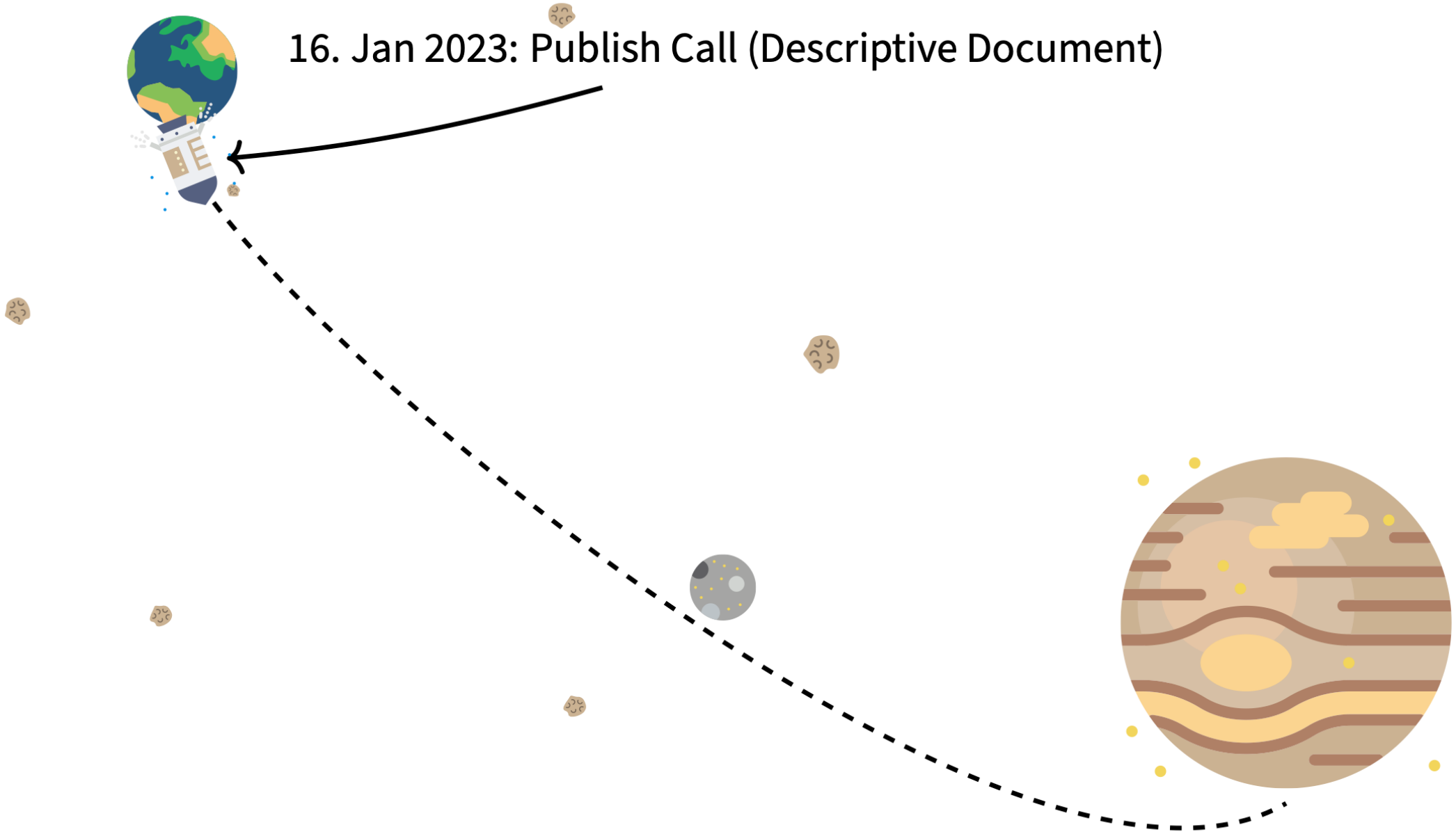
Hosting Entity Decision
15.06.2022

PREPARING FOR LAUNCH OF PROCUREMENT

- Mission planning
  - Preparing descriptions, conditions, requirements, evaluation
  - Regular meetings
  - Started already early in 2022
  - Location: **Earth**
- Target: **JUPITER**
  - Booster
  - Cluster
  - Storage
  - (Machine Hall)

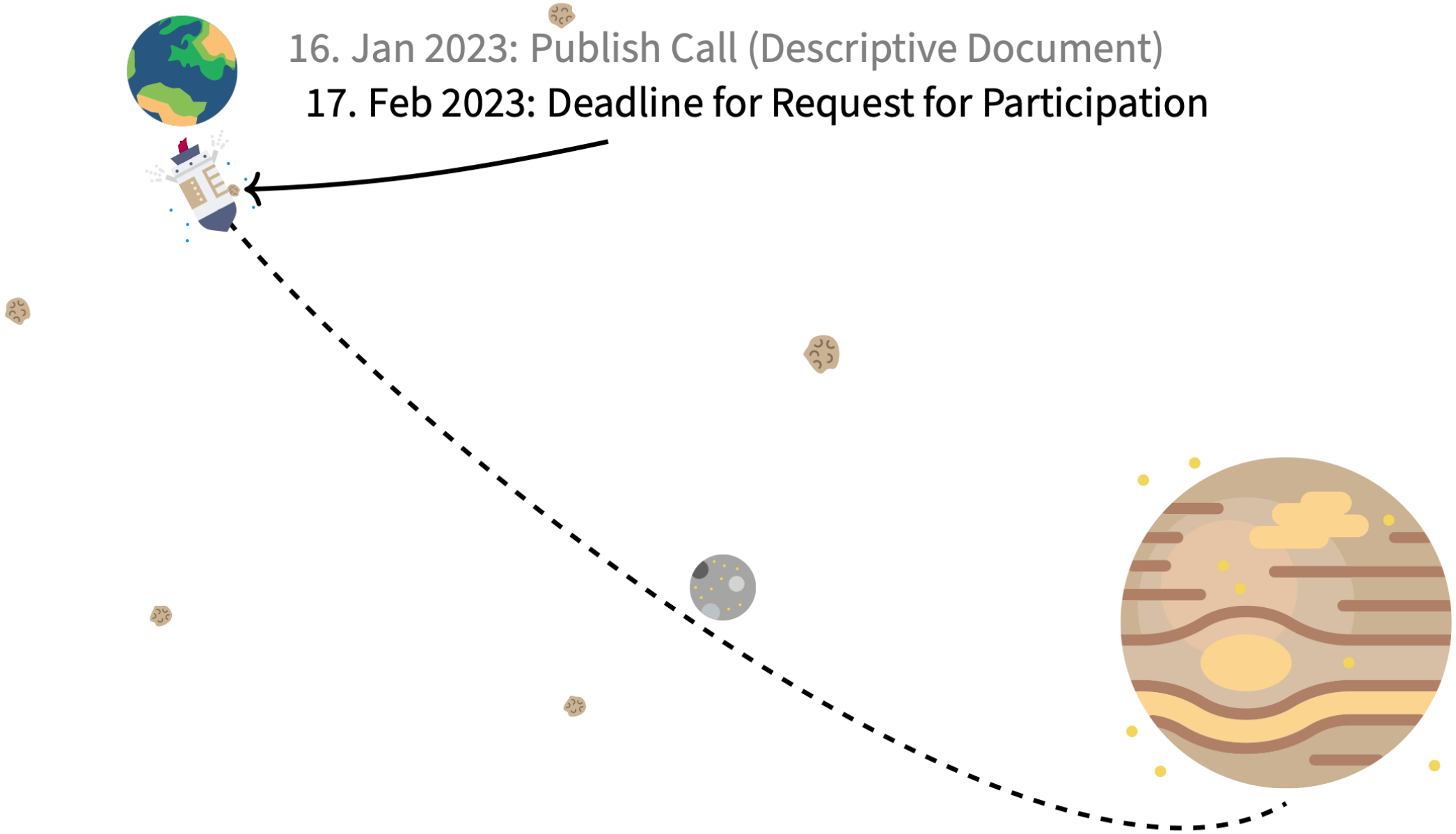


16. Jan 2023: Publish Call (Descriptive Document)



16. Jan 2023: Publish Call (Descriptive Document)

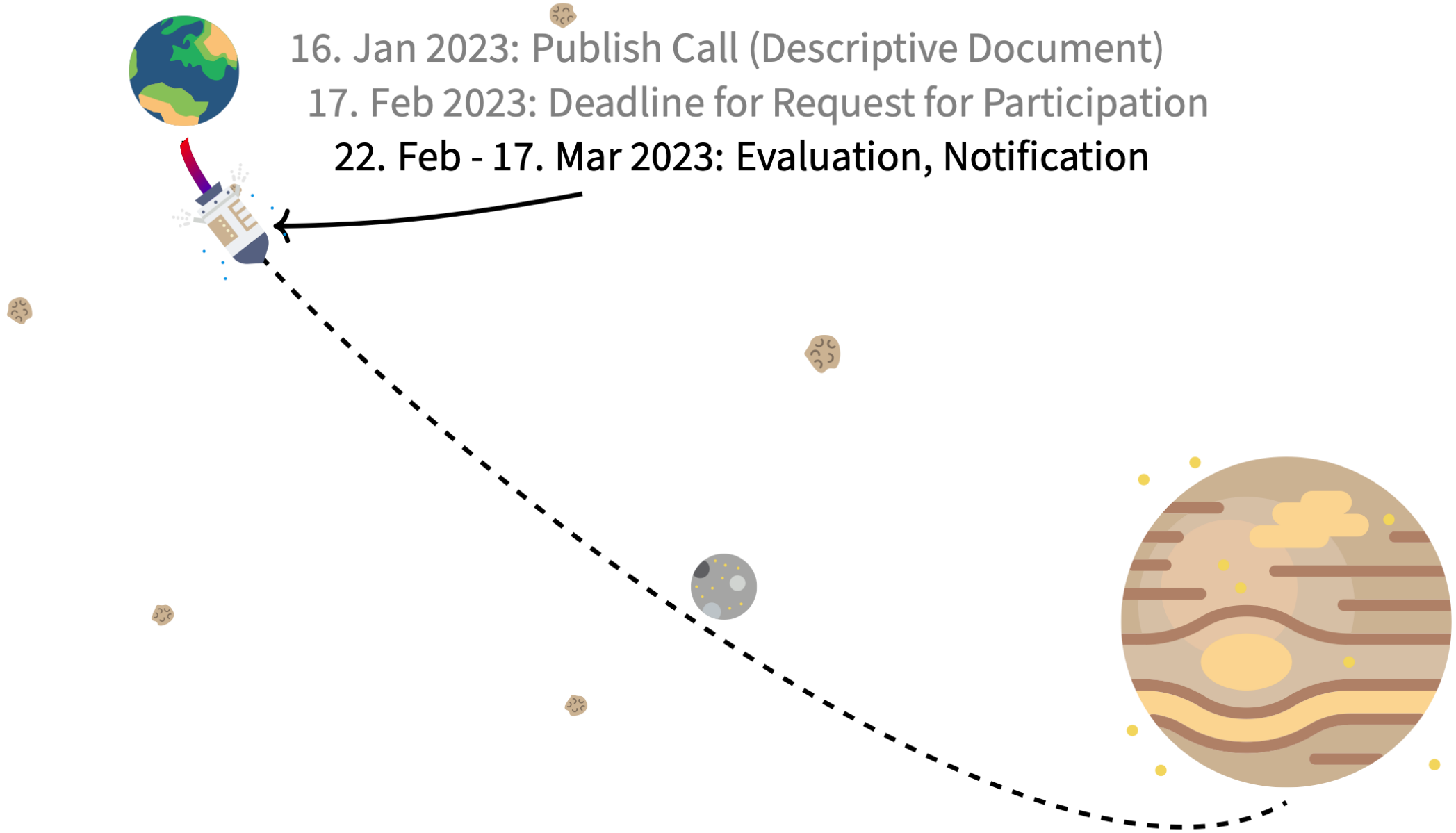
17. Feb 2023: Deadline for Request for Participation



16. Jan 2023: Publish Call (Descriptive Document)

17. Feb 2023: Deadline for Request for Participation

22. Feb - 17. Mar 2023: Evaluation, Notification

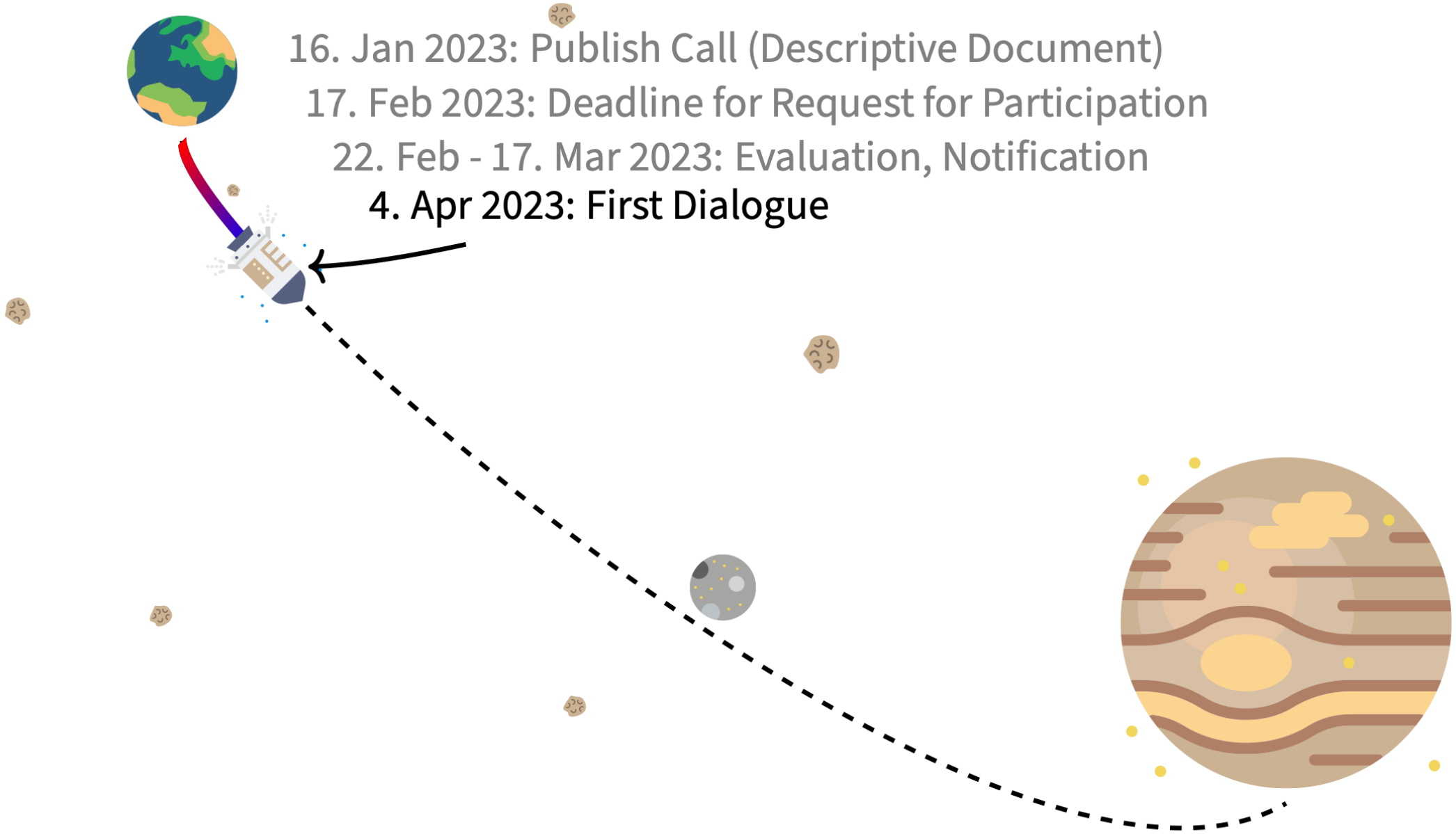


16. Jan 2023: Publish Call (Descriptive Document)

17. Feb 2023: Deadline for Request for Participation

22. Feb - 17. Mar 2023: Evaluation, Notification

4. Apr 2023: First Dialogue



16. Jan 2023: Publish Call (Descriptive Document)

17. Feb 2023: Deadline for Request for Participation

22. Feb - 17. Mar 2023: Evaluation, Notification

4. Apr 2023: First Dialogue

3.-5. May 2023: Second Dialogue

16. Jan 2023: Publish Call (Descriptive Document)

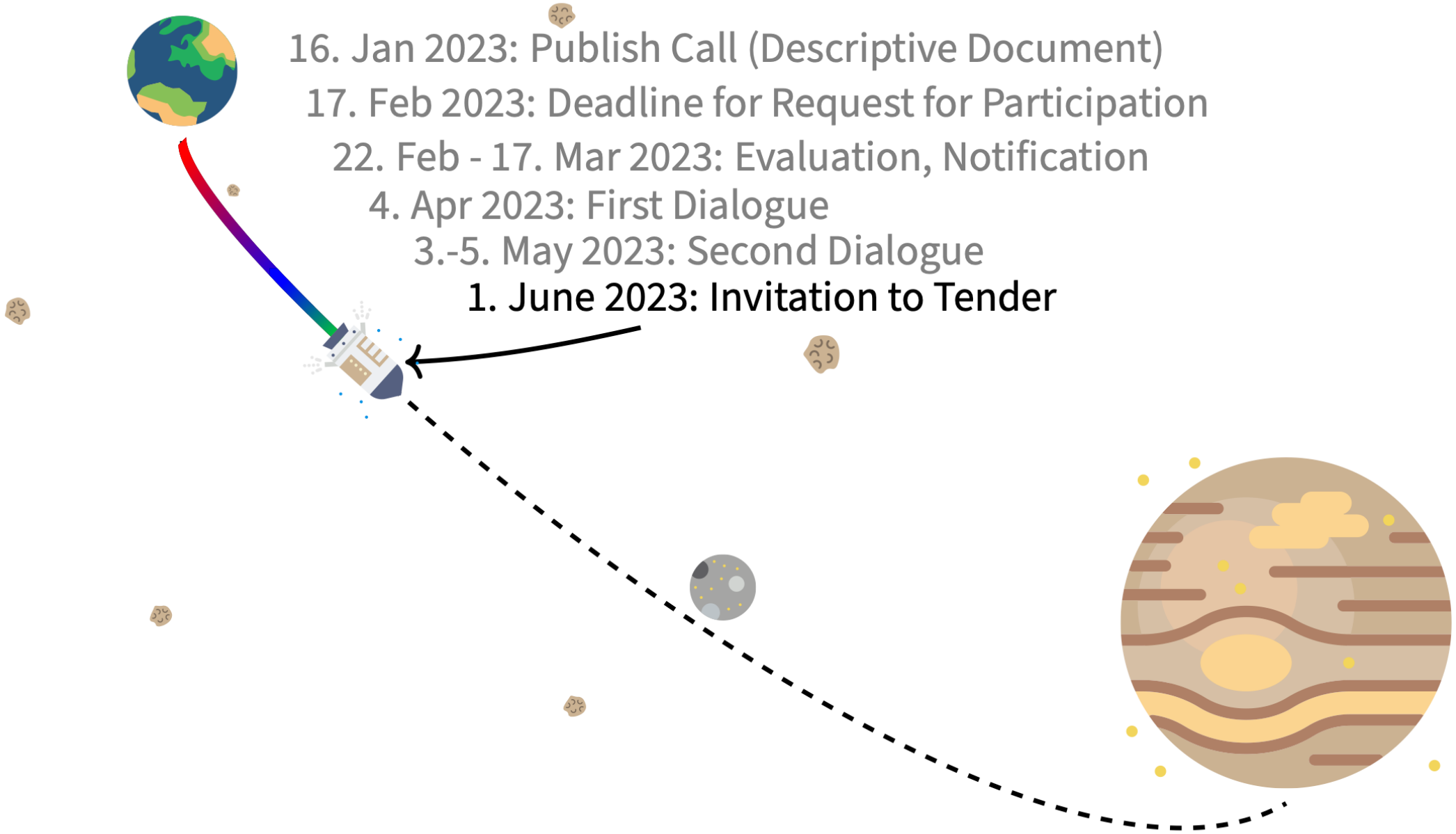
17. Feb 2023: Deadline for Request for Participation

22. Feb - 17. Mar 2023: Evaluation, Notification

4. Apr 2023: First Dialogue

3.-5. May 2023: Second Dialogue

1. June 2023: Invitation to Tender



16. Jan 2023: Publish Call (Descriptive Document)

17. Feb 2023: Deadline for Request for Participation

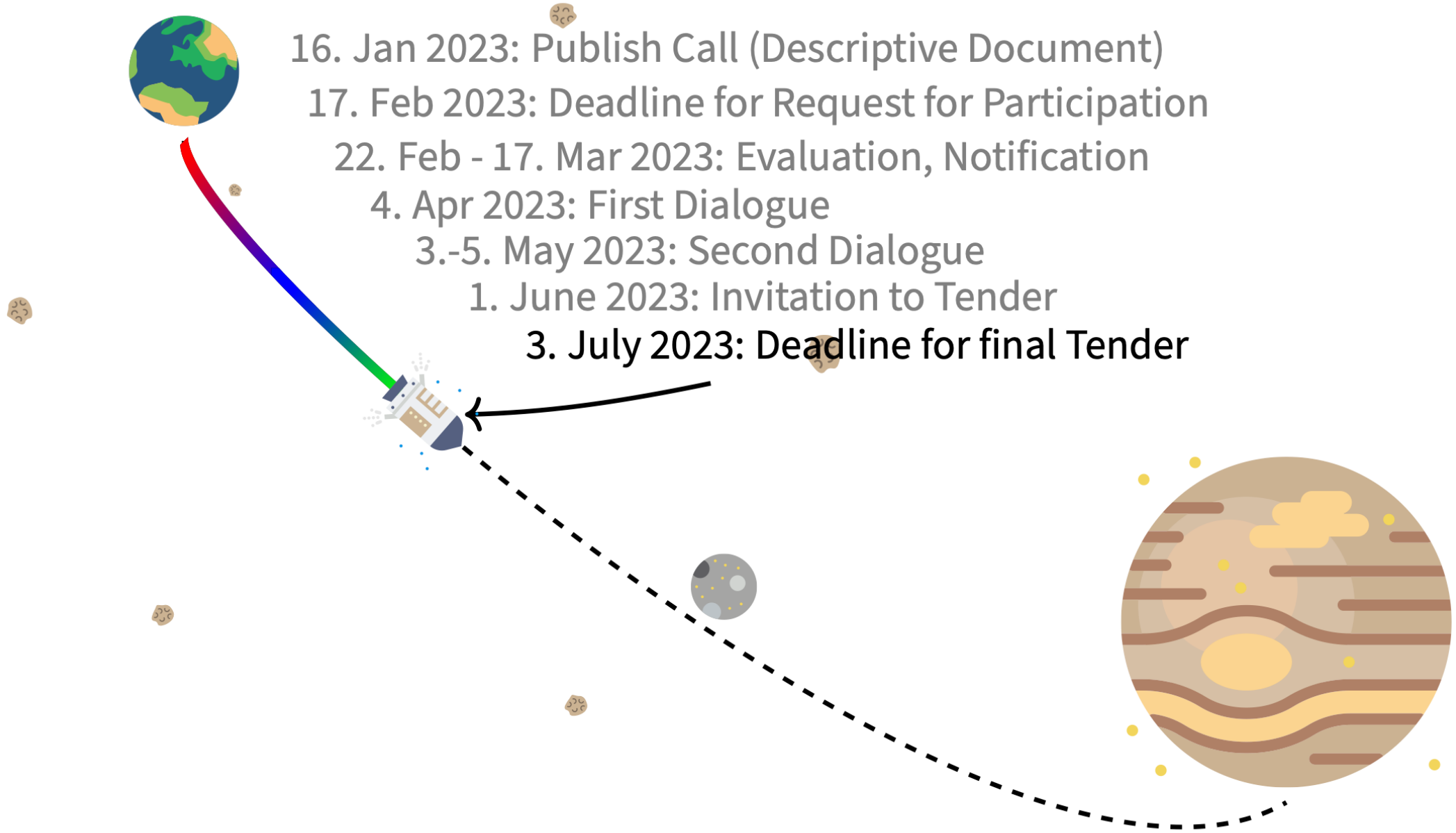
22. Feb - 17. Mar 2023: Evaluation, Notification

4. Apr 2023: First Dialogue

3.-5. May 2023: Second Dialogue

1. June 2023: Invitation to Tender

3. July 2023: Deadline for final Tender



16. Jan 2023: Publish Call (Descriptive Document)

17. Feb 2023: Deadline for Request for Participation

22. Feb - 17. Mar 2023: Evaluation, Notification

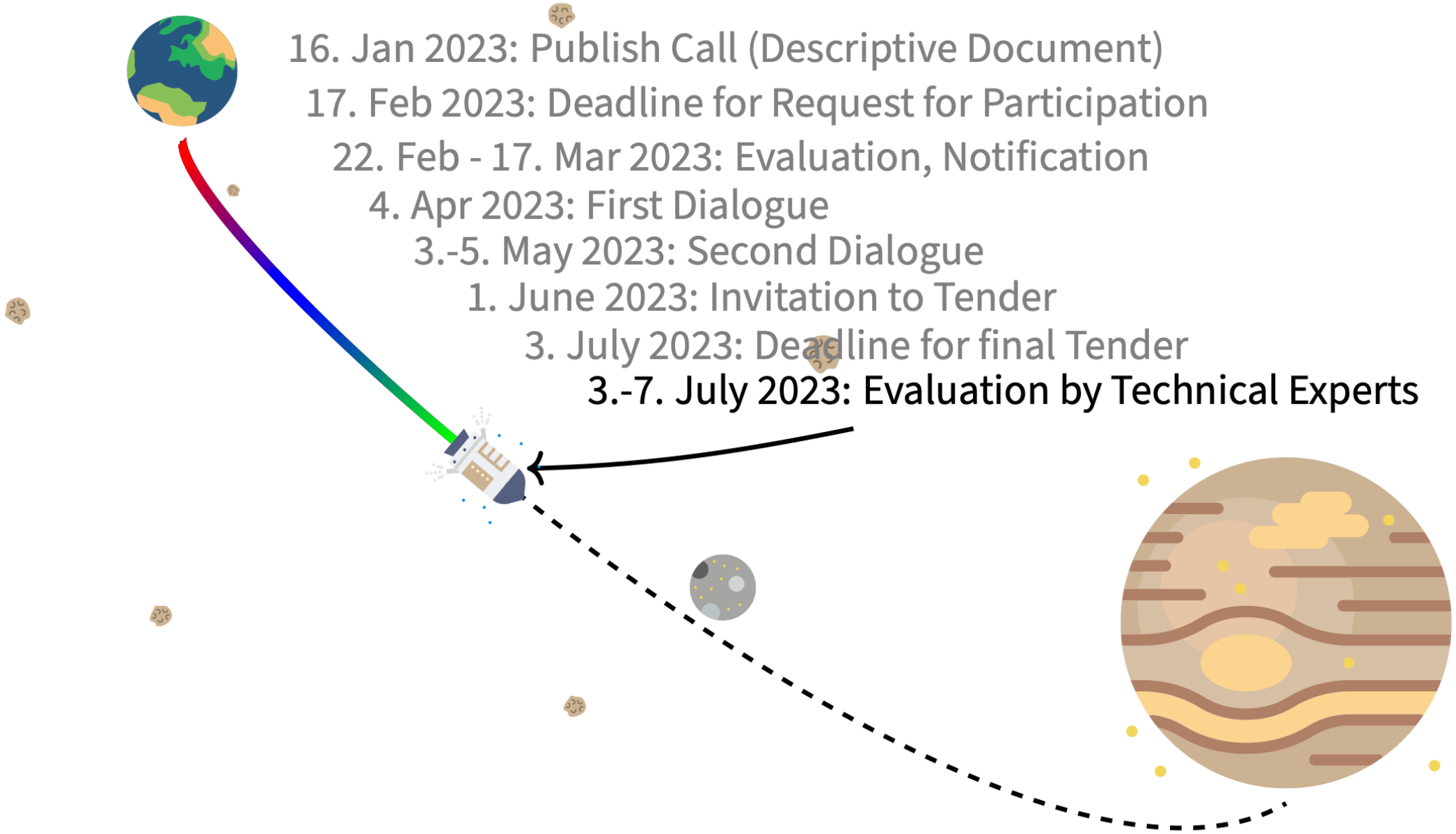
4. Apr 2023: First Dialogue

3.-5. May 2023: Second Dialogue

1. June 2023: Invitation to Tender

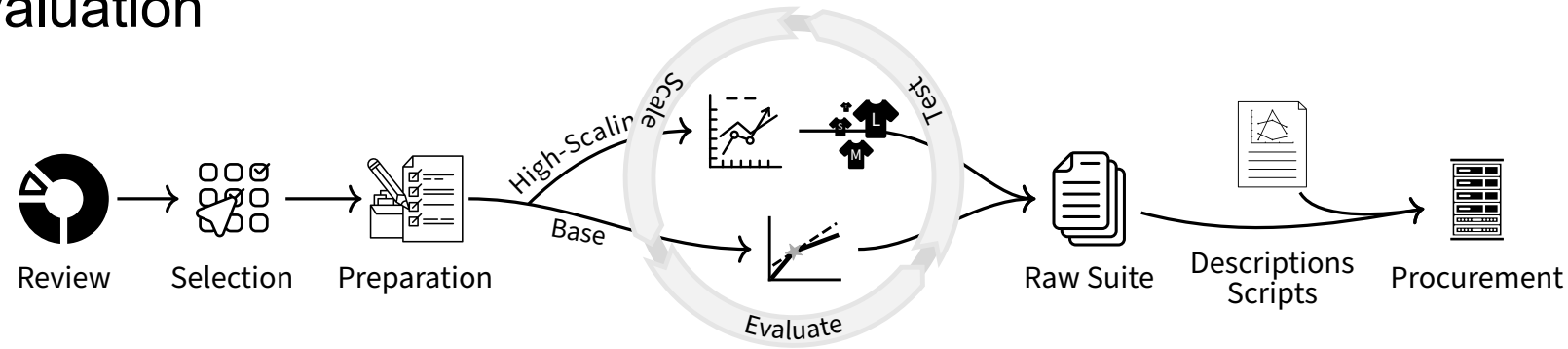
3. July 2023: Deadline for final Tender

3.-7. July 2023: Evaluation by Technical Experts

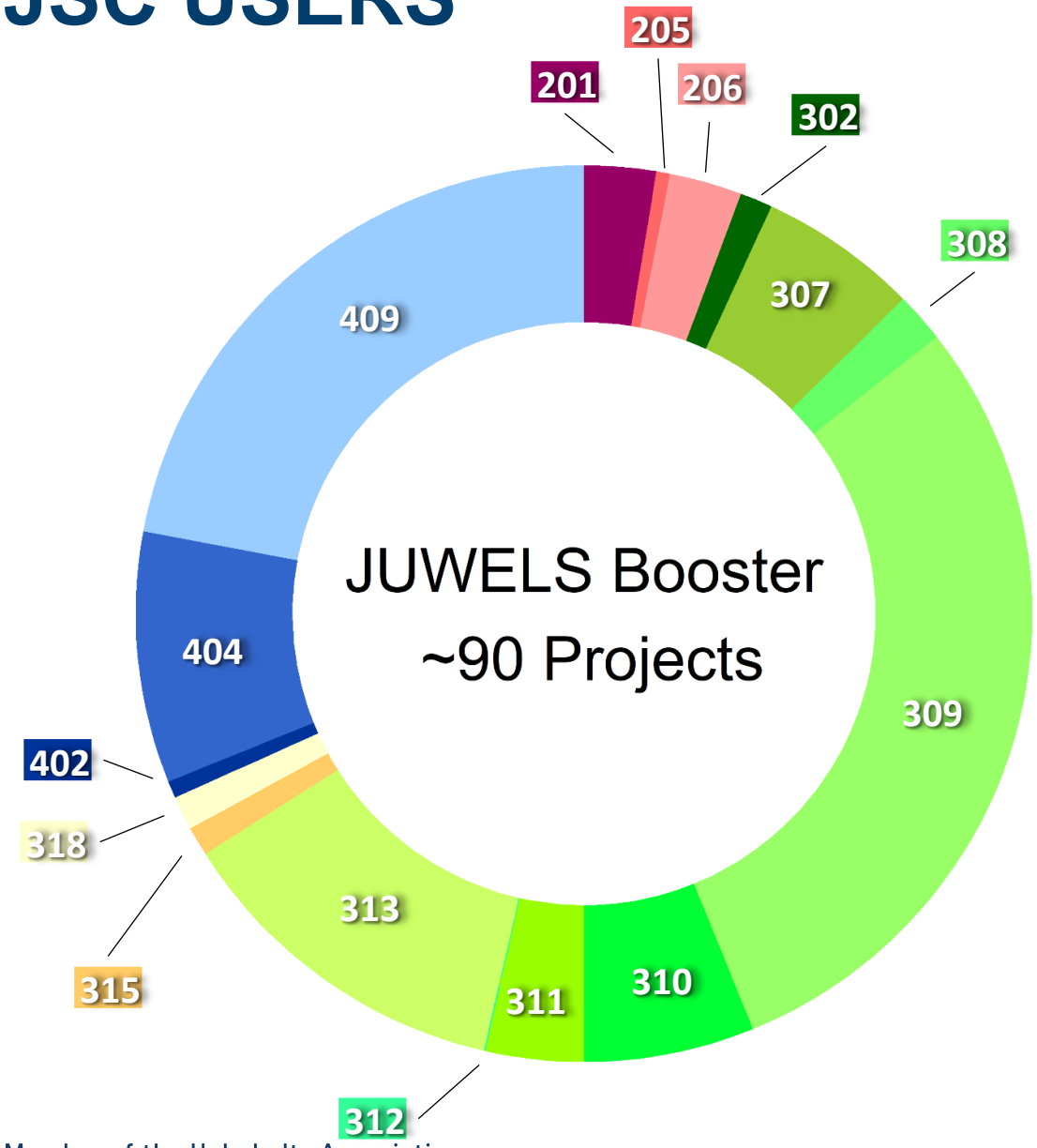


EVALUATION

- Criteria
  - Requirements to project planning, etc.
  - Technical requirements to overarching design and details
  - Performance of applications, benchmarks
    - Total cost of ownership (TCO): How much science for money
    - Further categories (Synthetic Benchmarks, High-Scaling Applications)
- Quantified evaluation



JSC USERS



Reseach Fields

- 201 Basic Biological and Medical Research
- 205 Medicine
- 206 Neurosciences
- 302 Chemical Solid State and Surface Research
- 307 Condensed Matter Physics
- 308 Optics, Quantum Optics and Physics of Atoms, Molecules and Plasmas
- 309 Particles, Nuclei and Fields
- 310 Statistical Physics, Soft Matter, Biological Physics, Nonlinear Dynamics
- 311 Astrophysics and Astronomy
- 312 Mathematics
- 313 Atmospheric Science, Oceanography and Climate Research
- 315 Geophysics and Geodesy
- 318 Water Research
- 402 Mechanics and Constructive Mechanical Engineering
- 404 Heat Energy Technology, Thermal Machines, Fluid Mechanics
- 409 Computer Science

→ Define Benchmarks

SUB-BENCHMARKS, VARIANTS

- Type of benchmarks
  - Applications benchmarks
  - Synthetic benchmarks
- Execution targets
  - JUPITER Booster (GPU, CPU)
  - JUPITER Cluster (CPU)
  - MSA
- Application benchmark categories
  - TCO
  - High-Scaling

		Booster			Cluster	MSA
Before Dialogue	After Dialogue	GPU	GPU High-Scale	CPU	CPU	
Amber	~~Amber~~	~~✓~~				
Arbor	Arbor	✓	✓			
Chroma	Chroma	✓	✓			
Gromacs	Gromacs (2)	✓				
ICON	ICON (2)	✓				
JUQCS	JUQCS	✓	✓			✓
nekRS	nekRS	✓	✓			
ParFlow	ParFlow	✓				
PIConGPU	PIConGPU	✓	✓			
Quantum ESPRESSO	Quantum ESPRESSO	✓				
SOMA	~~SOMA~~	~~✓~~				
AI-MMoCLIP	AI-MMoCLIP	✓				
AI-NLP	AI-NLP	✓				
AI-ResNet	~~AI-Resnet~~	~~✓~~				
dynQCD	dynQCD				✓	
NAStJA	NAStJA				✓	
Graph500	Graph500			✓		
HPCG	HPCG	✓			✓	
HPL	HPL	✓			✓	
IOR	IOR			✓	✓	
LinkTest	LinkTest			✓	✓	✓
Multi-Flow IP	~~Multi-Flow IP~~			~~✓~~		
OSU	OSU (2)	✓		✓	✓	
STREAM	STREAM	✓			✓	

TCO APPLICATIONS

Total Cost of Ownership

- Traditional benchmark category
- ***How much of benchmark suite can be run in lifetime of system?*** Also: energy
- Key: same metric for each benchmark
  - Unit: time / s
  - Needed to convert rate → time
- One reference run for formula (e.g. 8 nodes); additional strong-scaled runs (e.g. 4, 16)
- Weights per individual benchmark

HIGH-SCALING APPLICATIONS

- Novel category for us
- Give benchmarks a focus on large-scaleness of system
- Compare execution on full* JUWELS Booster to full* JUPITER Booster
  - *: Use 50 PFLOP/s$^{\text{th. peak}}$ part of JUWELS Booster

  → compare to 1000 PFLOP/s$^{\text{th. peak}}$ part of JUPITER Booster
- Challenging design, challenging commitments
  - Design for unknown system, unknown device, unknown memory size
    Introduce 3 memory variants: small ($^2/_4$), medium ($^3/_4$), high ($^4/_4$ JWB A100 memory)
  - Many tests on scale at JUWELS Booster
  - Some internal extrapolations to JUPITER scale

- Arbor
  tiny ($^1/_4$), small, medium, large
- Chroma
  small, medium, large
- JUQCS
  small, large
- nekRS
  small, medium, large
- PIConGPU
  small, medium, large

16. Jan 2023: Publish Call (Descriptive Document)

17. Feb 2023: Deadline for Request for Participation

22. Feb - 17. Mar 2023: Evaluation, Notification

4. Apr 2023: First Dialogue

3.-5. May 2023: Second Dialogue

1. June 2023: Invitation to Tender

3. July 2023: Deadline for final Tender

3.-7. July 2023: Evaluation by Technical Experts

until 20. Aug 2023: Governing Board Decision

16. Jan 2023: Publish Call (Descriptive Document)

17. Feb 2023: Deadline for Request for Participation

22. Feb - 17. Mar 2023: Evaluation, Notification

4. Apr 2023: First Dialogue

3.-5. May 2023: Second Dialogue

1. June 2023: Invitation to Tender

3. July 2023: Deadline for final Tender

3.-7. July 2023: Evaluation by Technical Experts

until 20. Aug 2023: Governing Board Decision

16. Jan 2023: Publish Call (Descriptive Document)

17. Feb 2023: Deadline for Request for Participation

22. Feb - 17. Mar 2023: Evaluation, Notification

4. Apr 2023: First Dialogue

3.-5. May 2023: Second Dialogue

1. June 2023: Invitation to Tender

3. July 2023: Deadline for final Tender

3.-7. July 2023: Evaluation by Technical Experts

until 20. Aug 2023: Governing Board Decision

23. Aug 2023:
Notification to Tenderers

12. Sep - 2. Oct 2023:
Contract Negotiations

16. Jan 2023: Publish Call (Descriptive Document)

17. Feb 2023: Deadline for Request for Participation

22. Feb - 17. Mar 2023: Evaluation, Notification

4. Apr 2023: First Dialogue

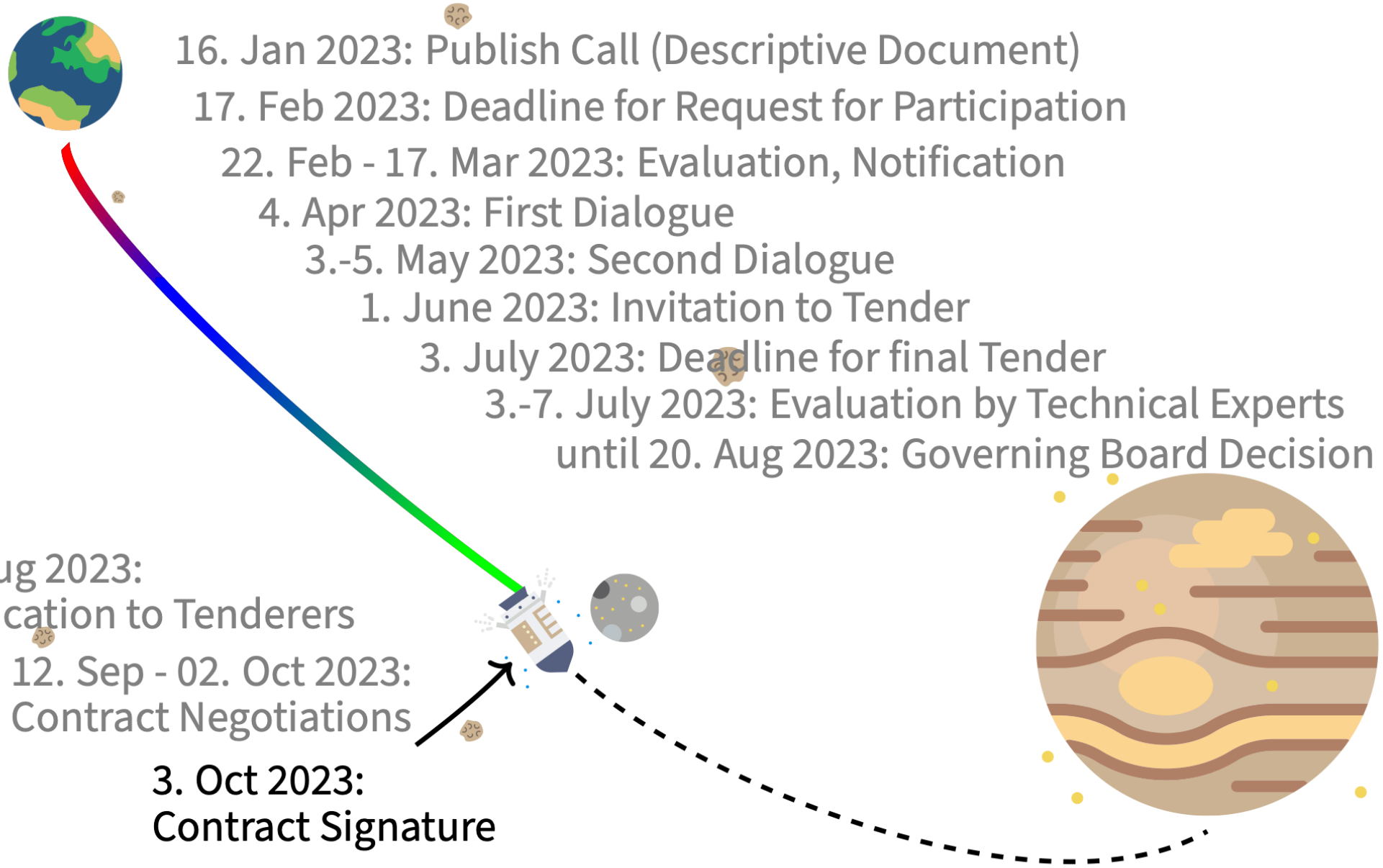
3.-5. May 2023: Second Dialogue

1. June 2023: Invitation to Tender

3. July 2023: Deadline for final Tender

3.-7. July 2023: Evaluation by Technical Experts

until 20. Aug 2023: Governing Board Decision

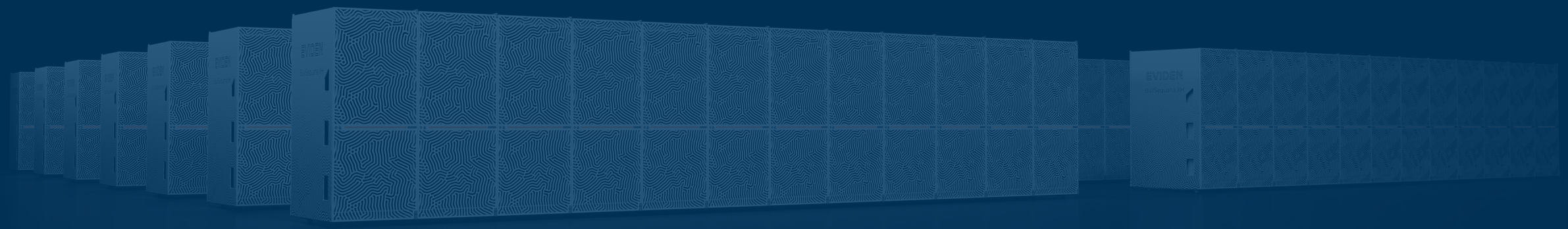


23. Aug 2023:
Notification to Tenderers

12. Sep - 02. Oct 2023:
Contract Negotiations

3. Oct 2023:
Contract Signature

PLANET FEATURE OBSERVATIONS



EuroHPC Joint Undertaking

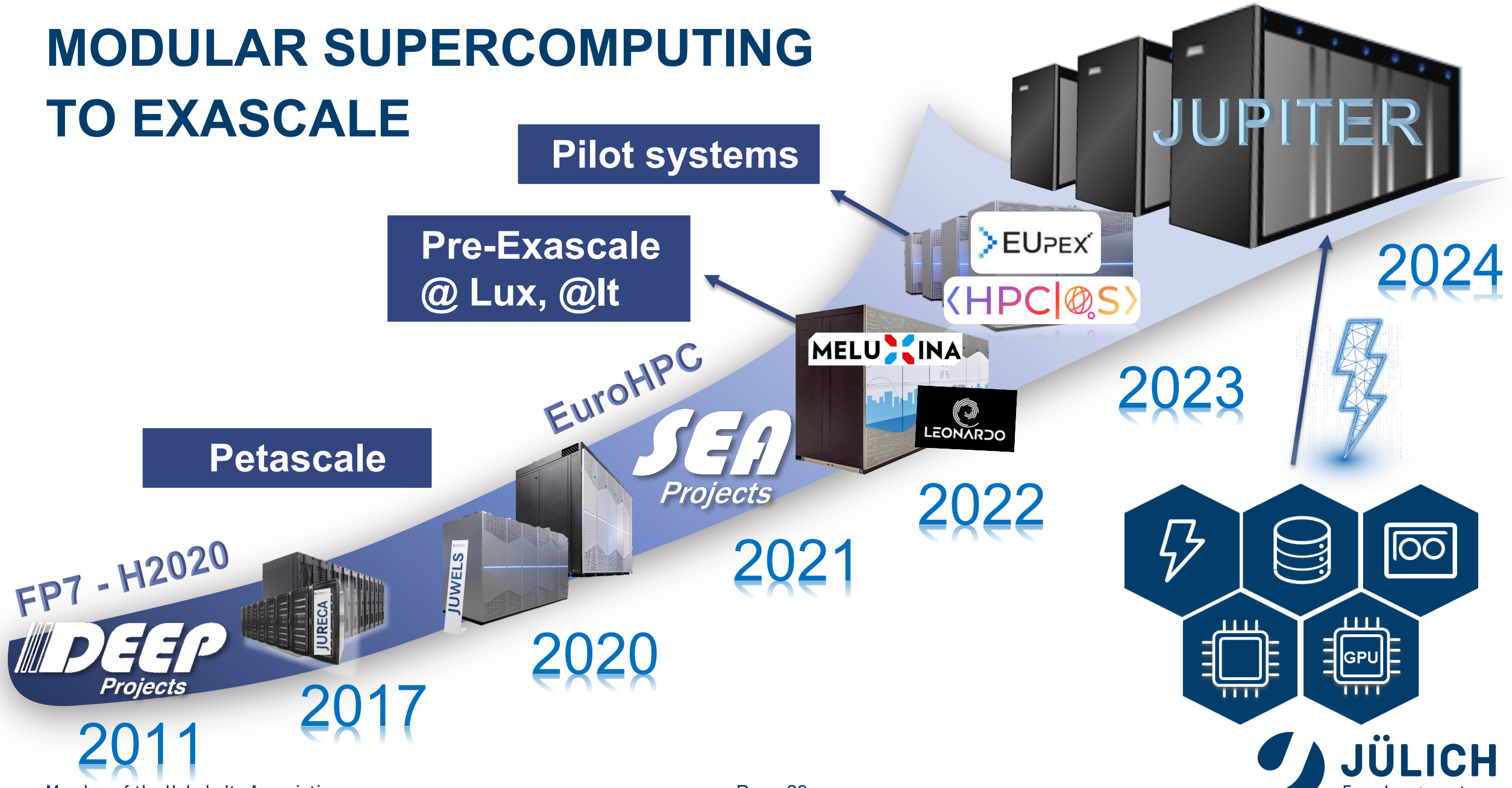
EVIDEN an atos business

JÜLICH Forschungszentrum

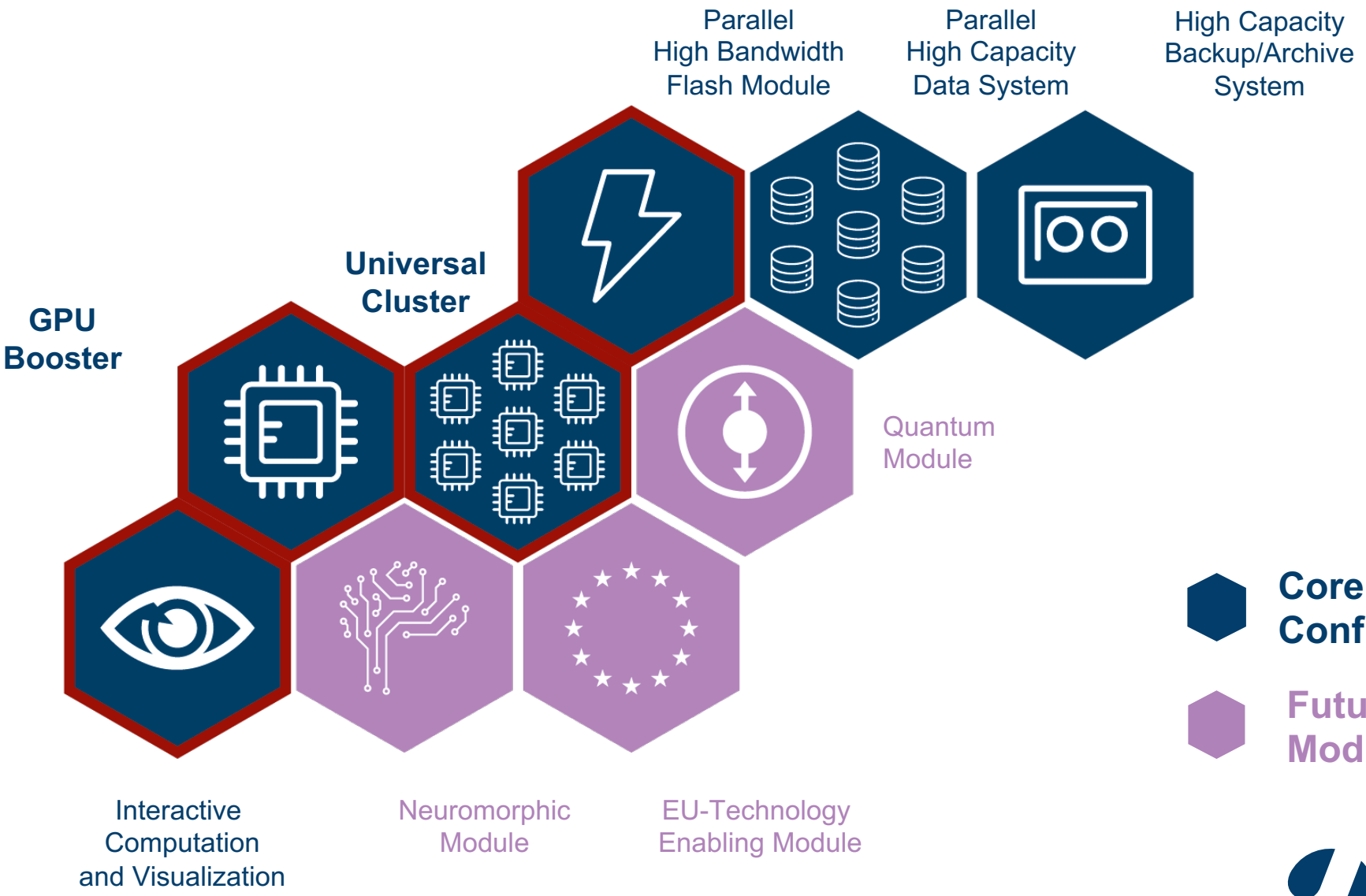
NVIDIA

ParTec MODULAR SUPERCOMPUTING

MODULAR SUPERCOMPUTING TO EXASCALE



JUPITER – HIGH-LEVEL ARCHITECTURE



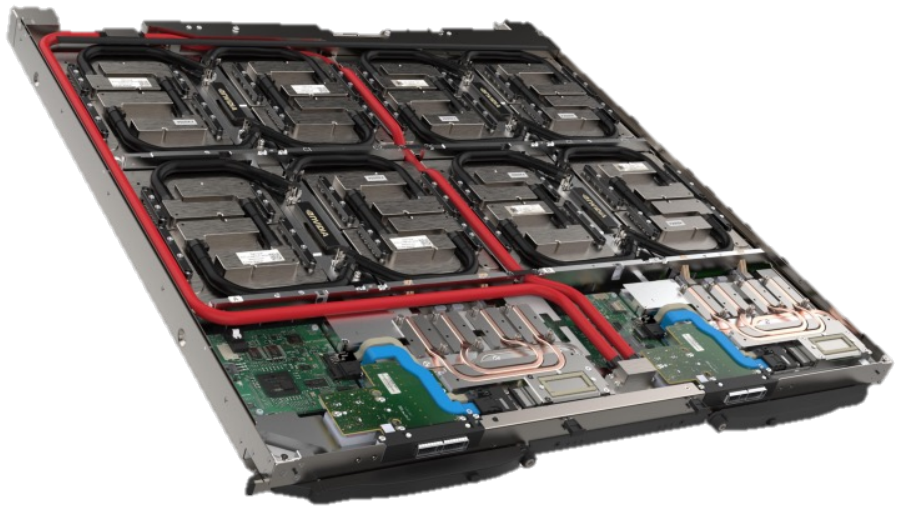
Core Configuration

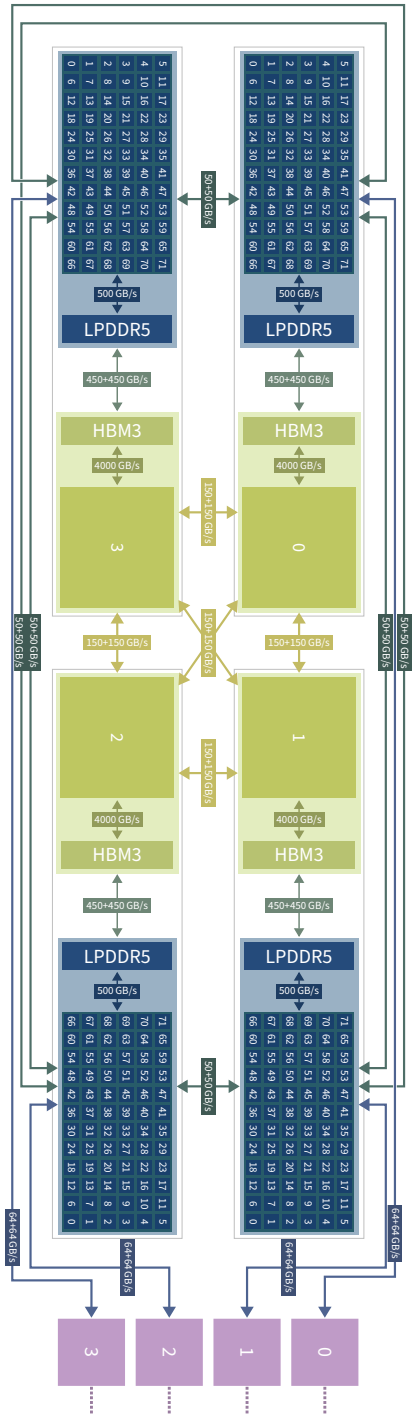
Future Call Modules

JUPITER – THE BOOSTER

Highly-Scalable Module for HPC and AI workloads

- **1** ExaFLOP/s (FP64, HPL)
- NVIDIA Grace-Hopper CG1
  - ~5900 compute nodes
  - **4×** CG1 chips per compute node
- NVIDIA Mellanox NDR
  - **4×** NDR200 NICs per compute node
- BullSequana XH3000
  - Direct Liquid Cooled blades
  - **2× compute node per blade**





BOOSTER NODE DESIGN

Node Specs

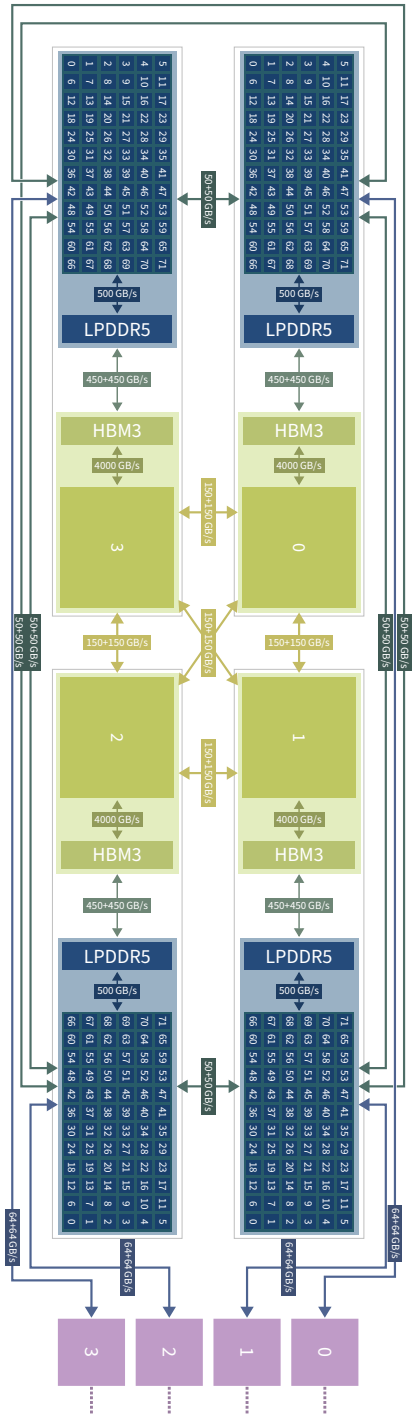
- 4× NVIDIA **Grace-Hopper** in SXM5 board (4× <750 W)
- 4× NVIDIA InfiniBand NDR200
- 480 GB LPDDR5X / 360 GB HBM3 (usable)
- NVLink 4
  - GPU-GPU 150 GB/s per dir
- Links: CPU-GPU 450 GB/s per dir, CPU-CPU 100 GB/s per dir

CPU Specs

- ARM Neoverse V2
  - SVE2/NEON (4x 128 bit vector op)
- 72 cores @ ~2.4GHz (~3.2 GHz turbo)
- 120 GB LPDDR5X (8 channels)
  - *≥450 GB/s*
  - *~150 ns latency*

GPU Specs

- H100
  - ~50 TFLOP/s (HPL single GPU)
- 90 GB HBM3
  - ≥3600 GB/s
  - ~450 ns latency

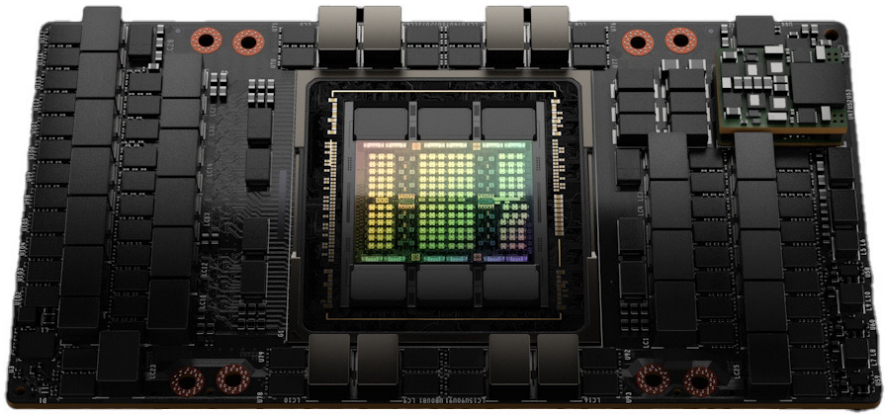


H100 GPU

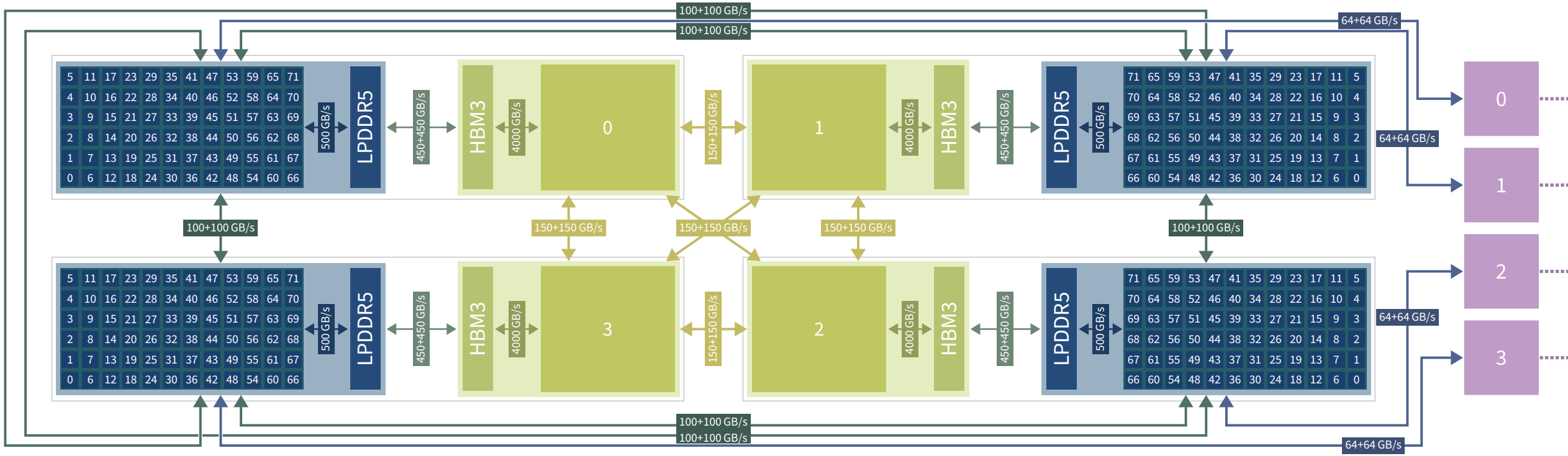
GH200 Flavor (not PCIe, not HBM2e, or HBM3e)

- 132 multiprocessors (*cores*)
  - 67 TFLOP/s $FP64^{TC}$
  - 989 TFLOP/s $FP16^{TC}$, 3958 TFLOP/s $FP8^{TC,Sp}$
- HBM3
  - 90 GB at 3.3 TB/s
- Notable new features
  - Thread-block clusters
  - Tensor memory accelerator
  - Transformer engine
- NVLink 4: 300 GB/s per GPU connection (900 GB/s)

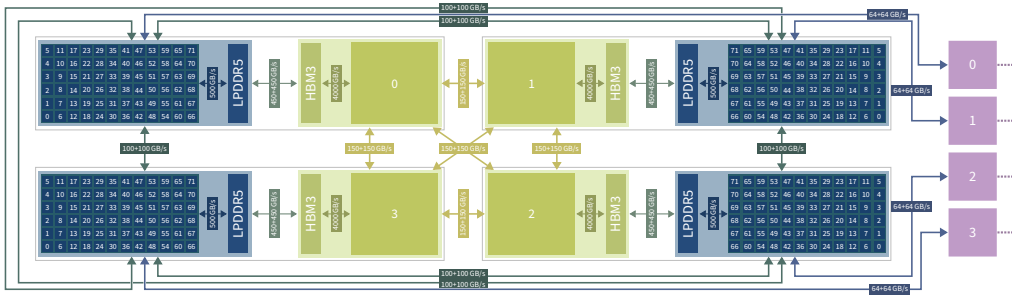
→ Whitepaper

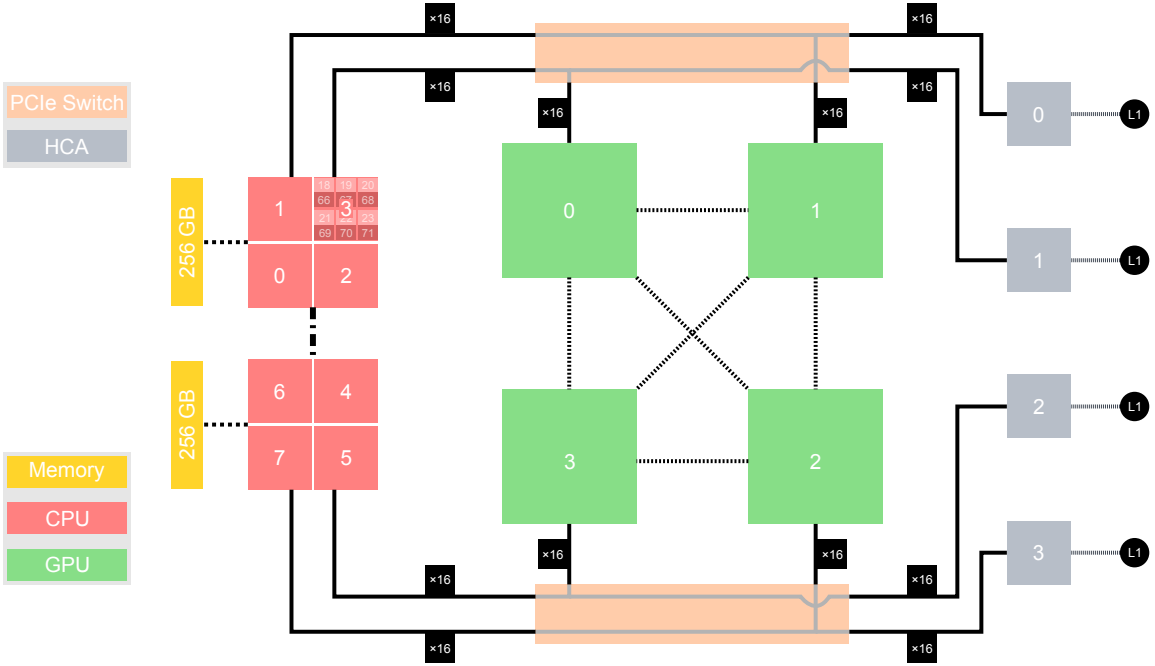


GRACE-HOPPER NODE OVERVIEW



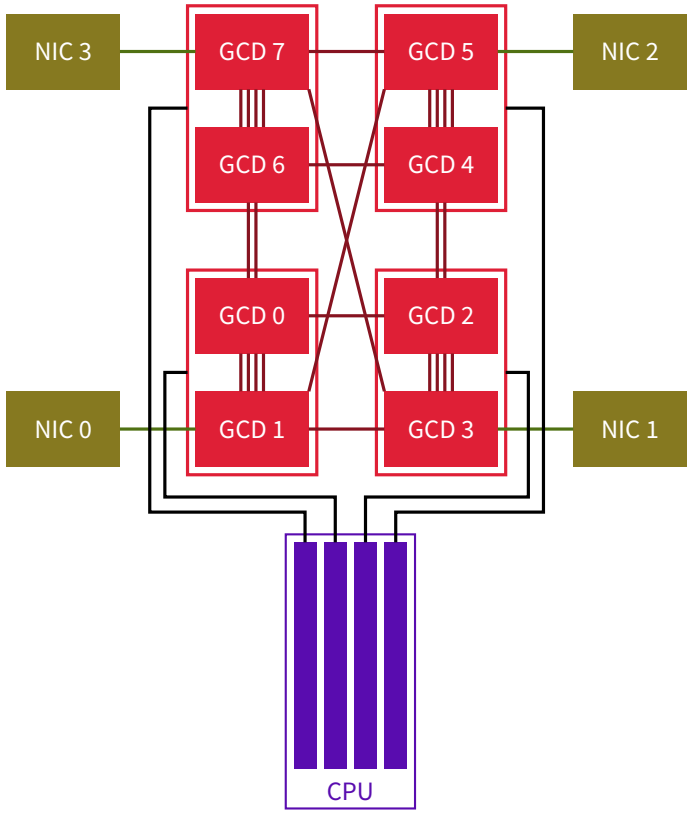
NODE COMPARISON



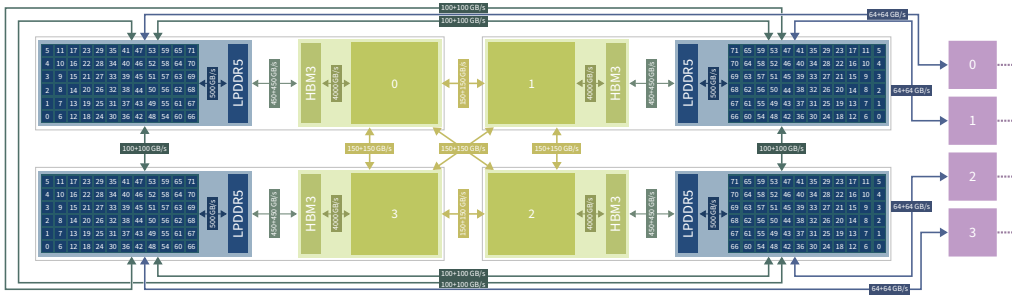


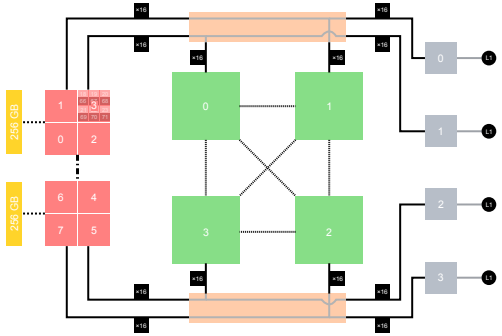
JUWELS Booster

NODE COMPARISON

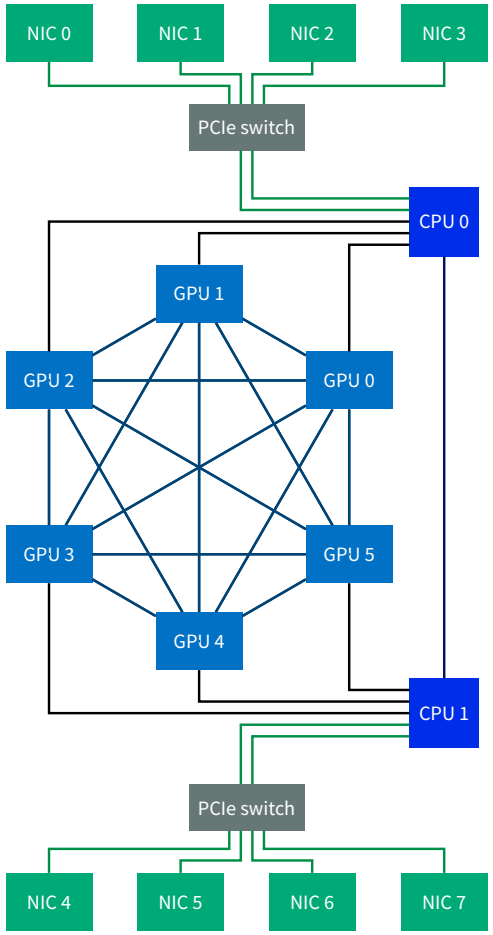


Frontier

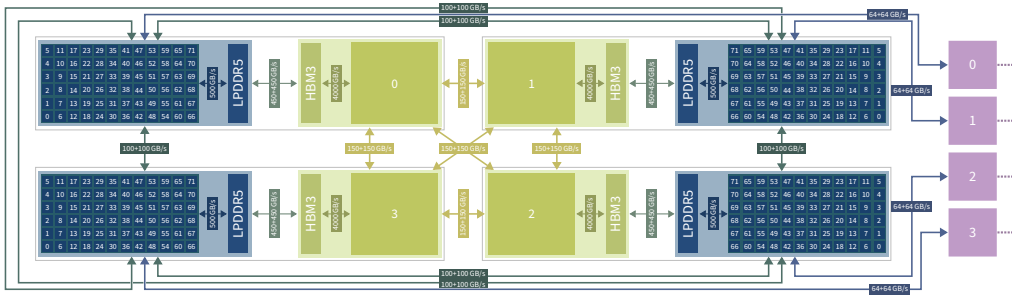


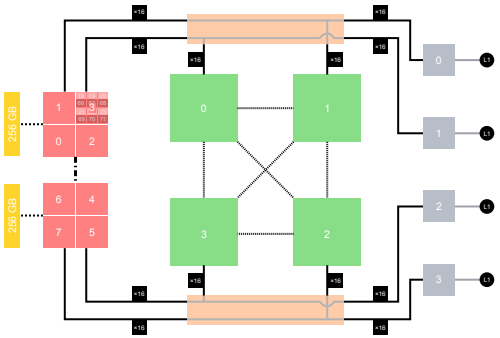


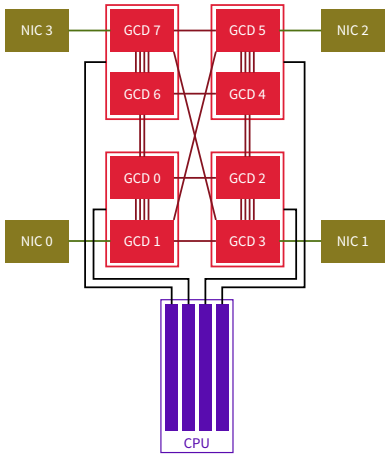
NODE COMPARISON



Aurora

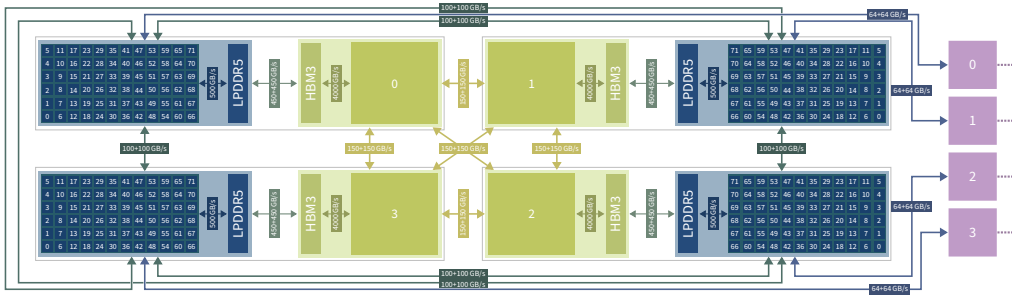


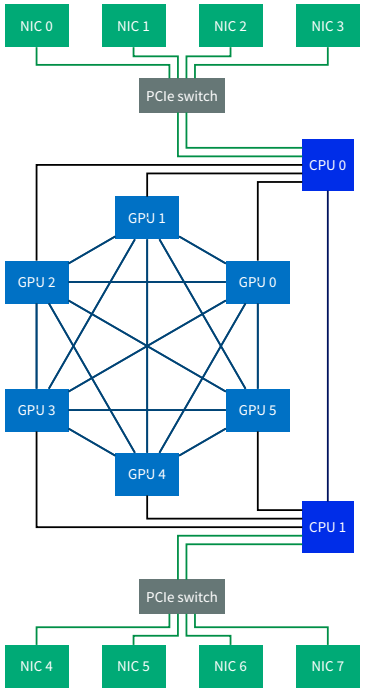


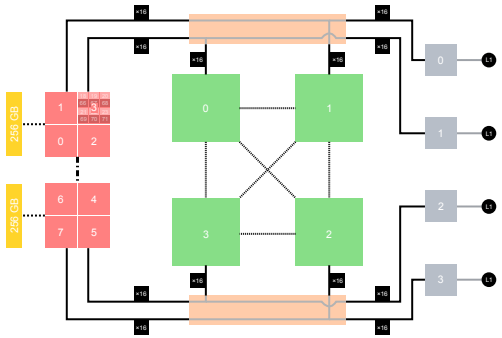


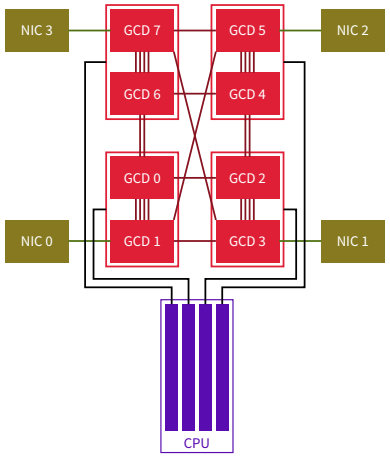
NODE COMPARISON

- JUWELS Booster: 2× CPU, 4× GPU, 4× IB
- JUPITER Booster: 4× CPU+GPU, 4× IB
- Frontier: 1× CPU, 4×(2× GPU), 4× Sling
- Aurora: 2× CPU, 6× GPU, 8× Sling
- El Capitan: APU





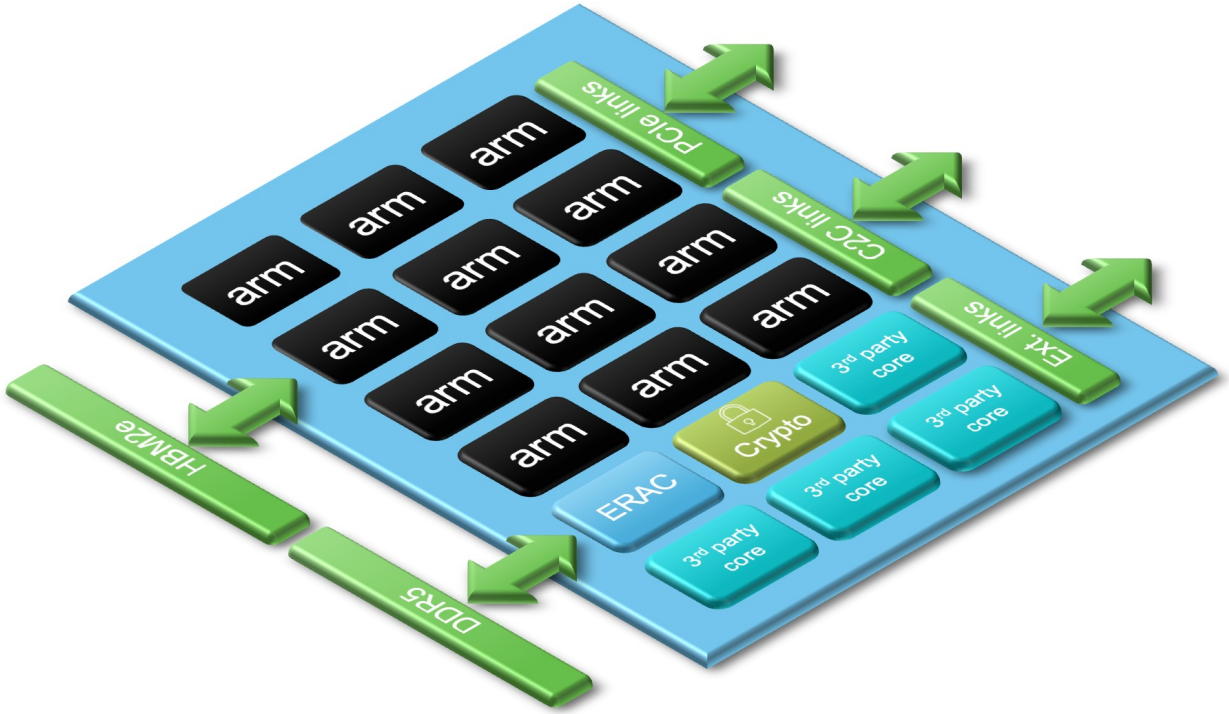




JUPITER – THE CLUSTER

General-Purpose Module for Mixed Workloads

- >5 PetaFLOP/s (FP64, HPL)
- SiPearl Rhea1
  - >1300 compute nodes
  - 2× CPUs per node
- NVIDIA Mellanox NDR
  - 1× NDR200 NICs per compute node
- BullSequana XH3000
  - Direct Liquid Cooled blades
  - 3× compute nodes per blade



JUPITER – STORAGE (SCRATCH)

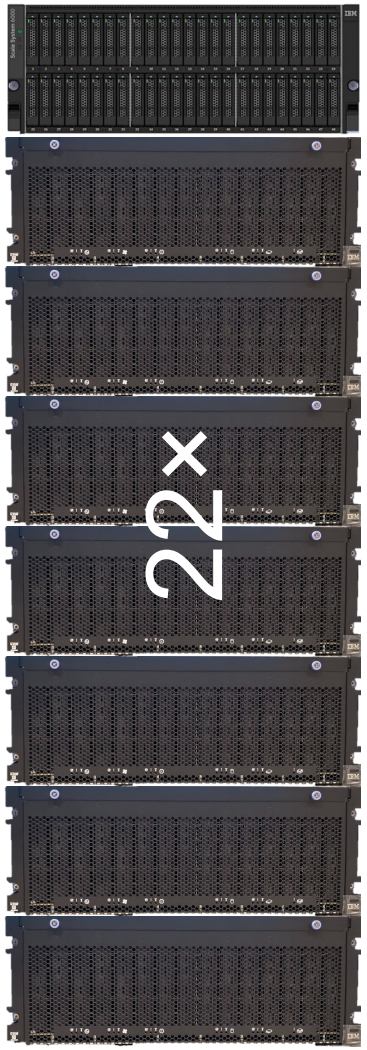
- Gross Capacity: 29 PB; Net Capacity: 21 PB
- Bandwidth: 2.1 TB/s Write, 3.1 TB/s Read
- 20× IBM SSS6000 Building Blocks (40 servers)
  - 2× NDR400 per server
  - 48× 30 TB NVMe drives per block
  - IBM Storage Scale (aka Spectrum Scale/GPFS)
- Manager and Datamover Nodes
- Exclusive for JUPITER
  - Integrated into InfiniBand fabric

Preliminary numbers, might change during installation

JUPITER – STORAGE (EXASTORE)

In kind contribution from JSC, not part of the JUPITER procurement

- Gross Capacity: 308 PB; Net Capacity: 210 PB
- Bandwidth: 1.1 TB/s Write, 1.4 TB/s Read
- 22× IBM SSS6000 Building Blocks (44 servers)
  - 2× NDR200 per server
  - 7× JBOD enclosures, each with 91x 22 TB Spinning Disks per block
  - IBM Storage Scale (aka Spectrum Scale/GPFS)
- Manager and Datamover Nodes
- Exclusive for JUPITER
  - Integrated into InfiniBand fabric

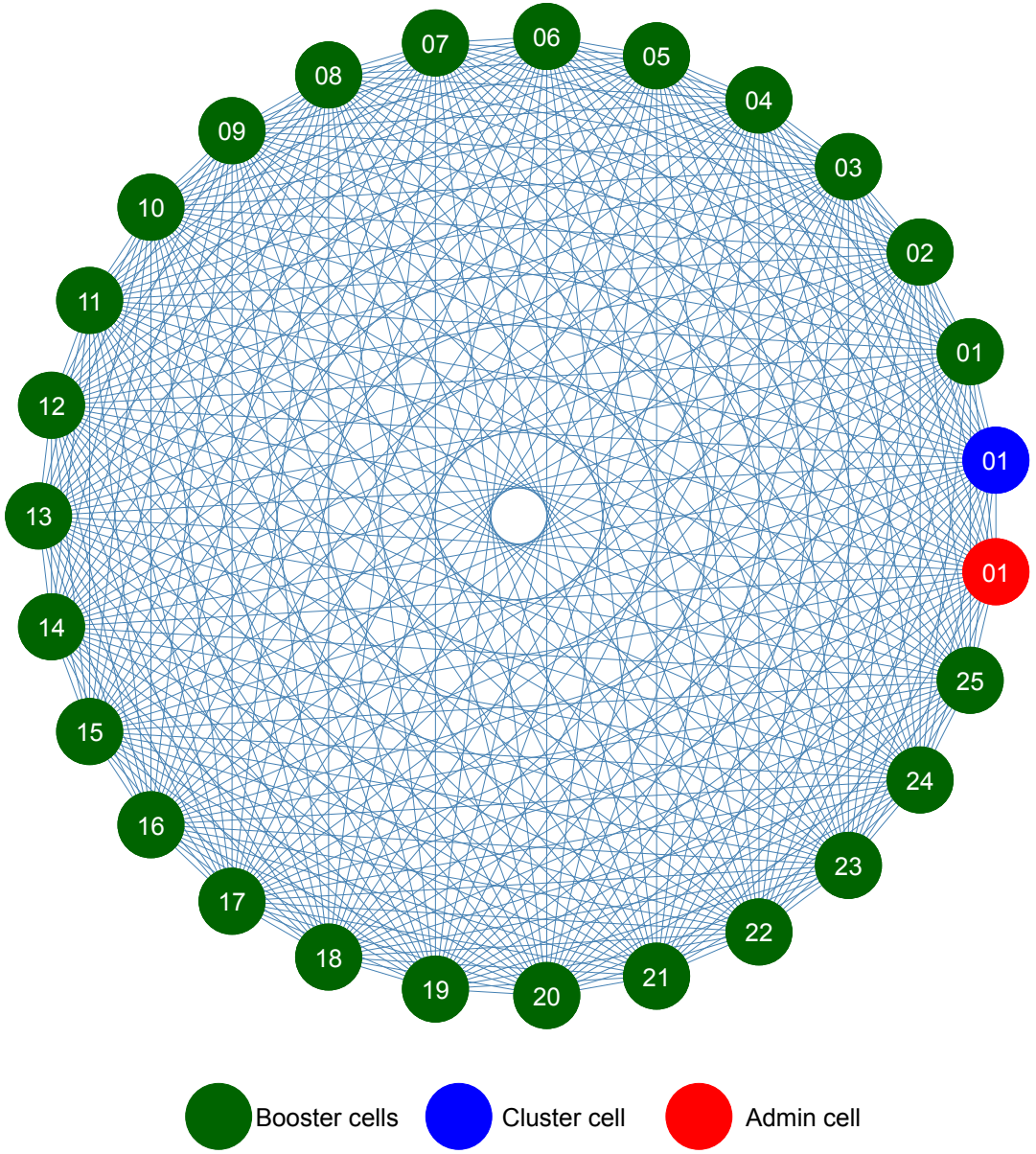


Preliminary numbers, might change during installation

JUPITER – INTERCONNECT

One Network to Rule Them All

- DragonFly+ topology, 27 cells in total
  - 25 Booster cells
  - 1 Cluster cell
  - 1 Admin cell
- Full fat tree inside each cell
- InfiniBand NDR 200 Gb/s



EVIDEN an atos business

NVIDIA

JÜLICH Forschungszentrum

JUPITER – INTERCONNECT

The Big Red Spot?

- Cluster module switch
- Booster module switch
- Booster node
- Flash storage node
- Misc node
- NDR link
- Admin module switch
- Cluster node
- Gateway node
- Storage node
- NDR200 link

JUPITER SYSTEM MANAGEMENT

”Power is nothing without control"

- Eviden SMC xScale
- ParaStation Modulo
  - Resource management
  - ParaStation MPI
  - HealthChecking/Ticket handling
- Ansible as provisioning system
- Scheduler: Slurm
- Software management: EasyBuild
- RedHat Enterprise Linux 9

LANDING SITE

LANDING SITE

LANDING IN DATA CENTER FOR JUPITER

- Current data center at limits (cooling, power)
- Original plan: new “traditional” data center
  - Time constraints
  - Exploding costs

- New plan: tailor-made **modular** (containerized) **data center** sized for JUPITER
  - Each container: own peripheral infrastructure
  - Connect pairs of containers for compute
  - Additional auxiliary containers

MODULAR DATA CENTER PRELIMINARY RENDERS

MODULAR DATA CENTER PRELIMINARY RENDERS

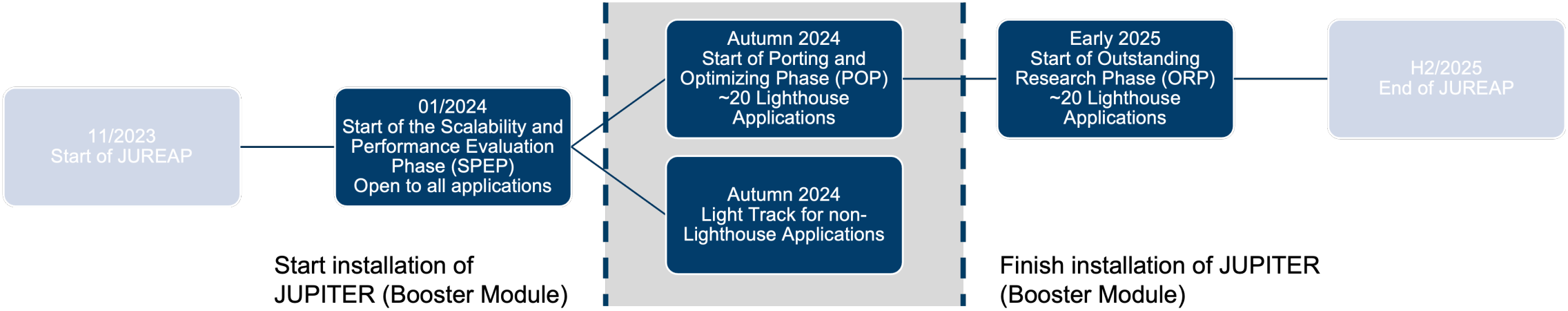

SCHLUM
TOI TOI
TOI TOI

PREPARING FOR DESCENT

JUREAP

JUPITER Research and Early Access Program

IMPORTANT: Timeline not final yet!



Phase 1: Scalability and Performance Evaluation Phase (SPEP)
Phase 2: Porting and Optimizing Phase (POP)
Phase 3: Outstanding Research Phase (ORP)

JUREAP call open since Jan‘2024
https://events.hifis.net/event/1239/

JUREAP

Seeding Exascale in Europe!

See JLESC poster!

jureap@fz-juelich.de • https://www.fz-juelich.de/en/ias/jsc/jupiter/jureap

JUPITER Research and Early Access Program

JÜLICH Forschungszentrum

Mitglied der Helmholtz-Gemeinschaft

Foto von Robert Wiedemann auf Unsplash

JUPITER TEST RACK

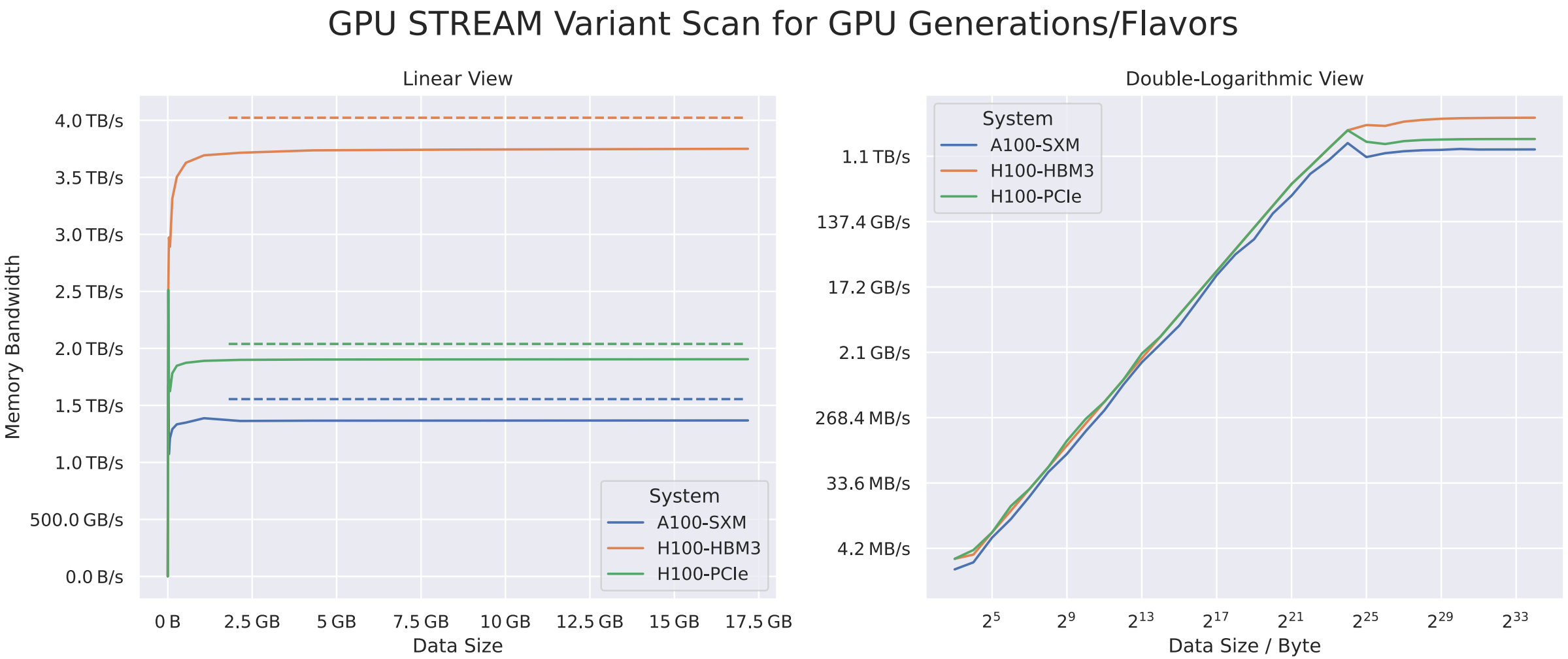
- Currently being installed
- Actual JUPITER node design
  - 2× nodes per blade
  - 4× Grace-Hopper per node
- Usage
  - System management preparations
  - JUREAP
  - Application porting
- Also: GH200 COTS test nodes

FIRST TRAINING RESULTS

GH200 COTS TEST NODES

- GH200 Prototype, commercial off-the-shelf hardware
- 2× test node, each with 1 Grace-Hopper
  - 1 Grace CPU (72 cores), 480 GB LPDDR5X RAM
  - 1 H100 GPU
  - TDP 700-1000 W
- Slightly different variant compared to JUPITER node design

GPU STREAM

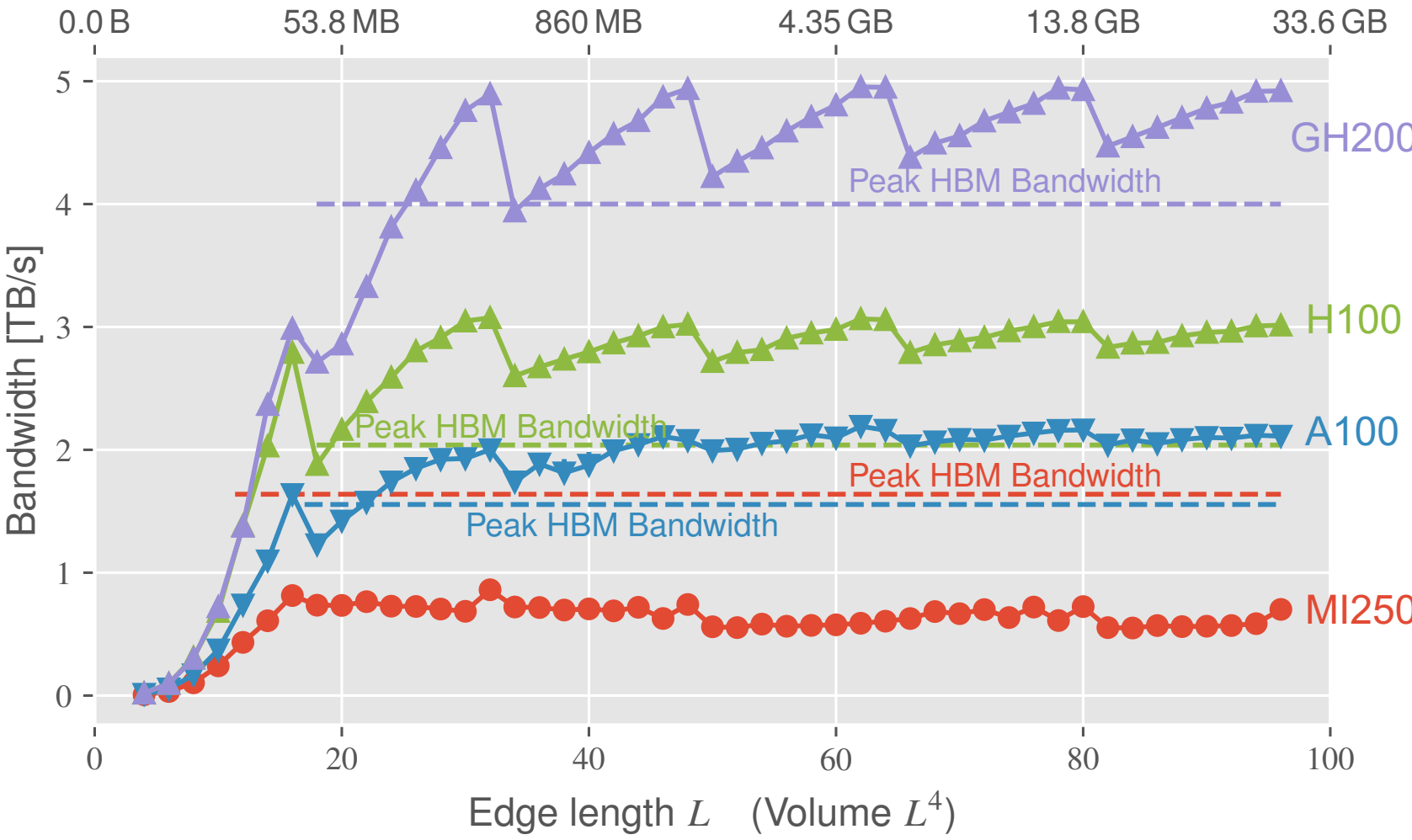


https://github.com/AndiH/CUDA-Cpp-STREAM

KOKKOS LQCD BENCHMARK

By Simon Schlepphorst / JSC

- Simplified Staggered Fermion Dirac Operator
- MI250
- A100
- H100 (PCIe)
- H100 (GH200)

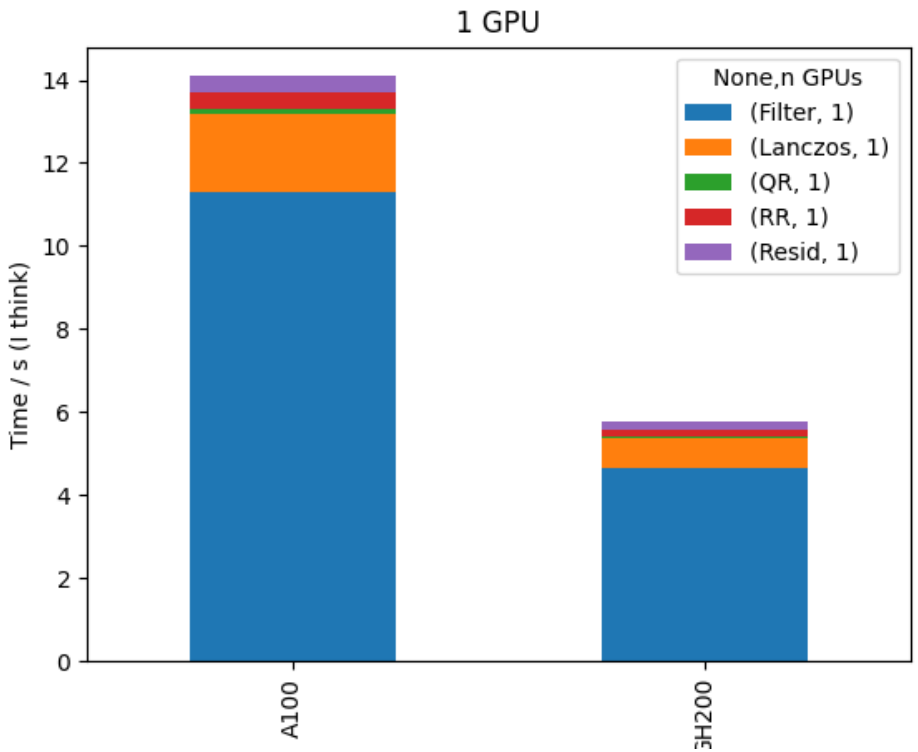


https://gitlab.jsc.fz-juelich.de/sdl-nqft/kokkos-lqcd-benchmarks

CHASE

By Xinzhe Wu / JSC

- Eigenvalue solver
- Benchmark: FP64, extremal eigenpairs of dense matrix (size 38000)



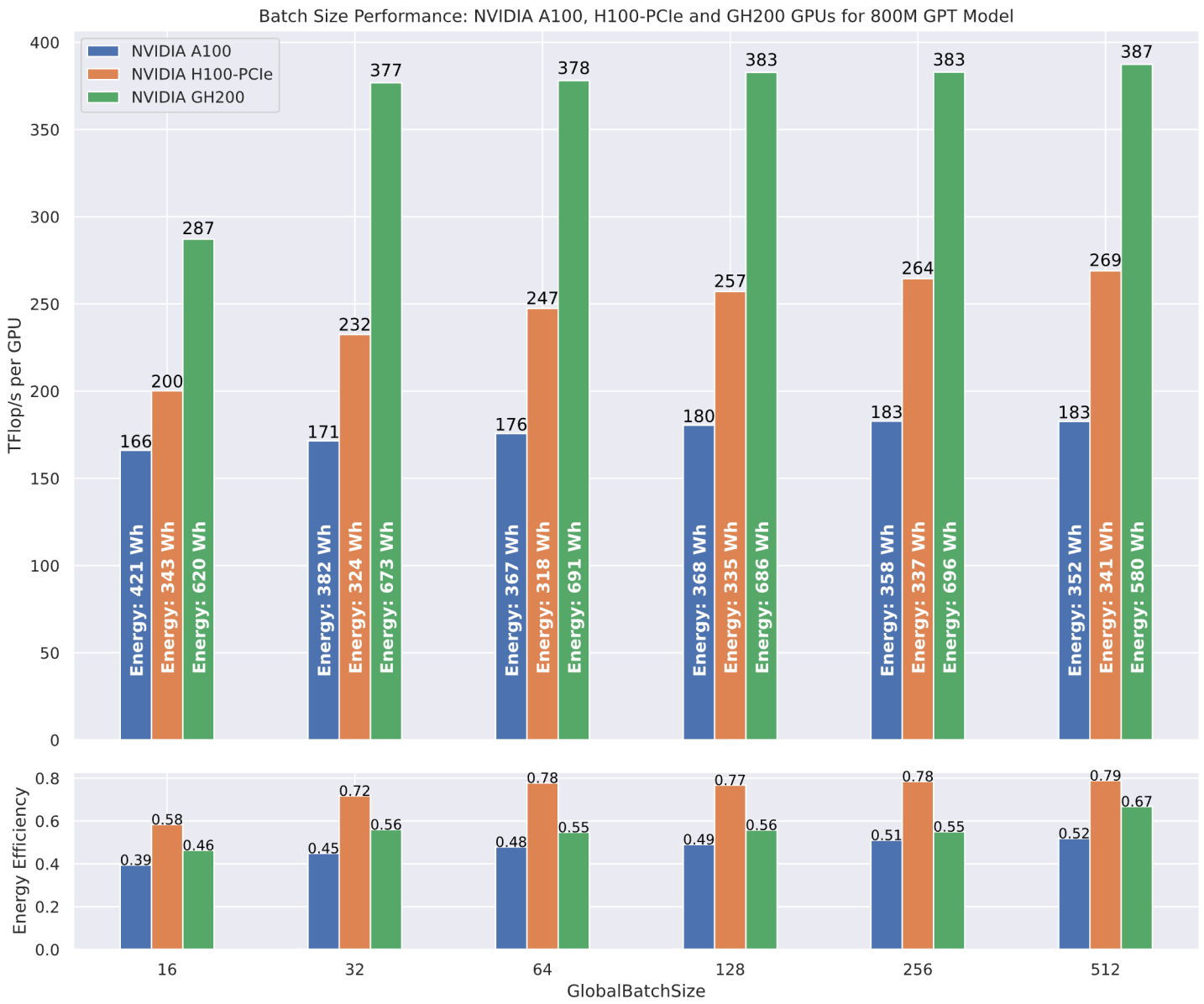
	1 node			2 nodes		
	JW-Booster (1 GPU/node)	GH200	speedup	JW-Booster (1 GPU/node)	GH200	speedup
All (s)	15.97	6.03	2.65×	13.64	4.97	2.74×
Lanczos(s)	1.89	0.72	2.62×	1.03	0.38	2.71×
Filter(s)	11.29	4.67	2.42×	10.67	3.88	2.75×
RR(s)	0.41	0.18	2.28×	0.46	0.27	1.7×
Resid(s)	0.38	0.17	2.24×	0.43	0.26	1.65×
QR(s)	0.12	0.018	6.67×	0.096	0.015	6.4×

https://github.com/ChASE-library/ChASE

MEGATRON

By Chelsea John / JSC

- Training of 800M GPT Model
- Increasing Batchsizes
- Up: Performance (FLOP/s)
  Down: Energy Efficiency (FLOP/s / Wh)

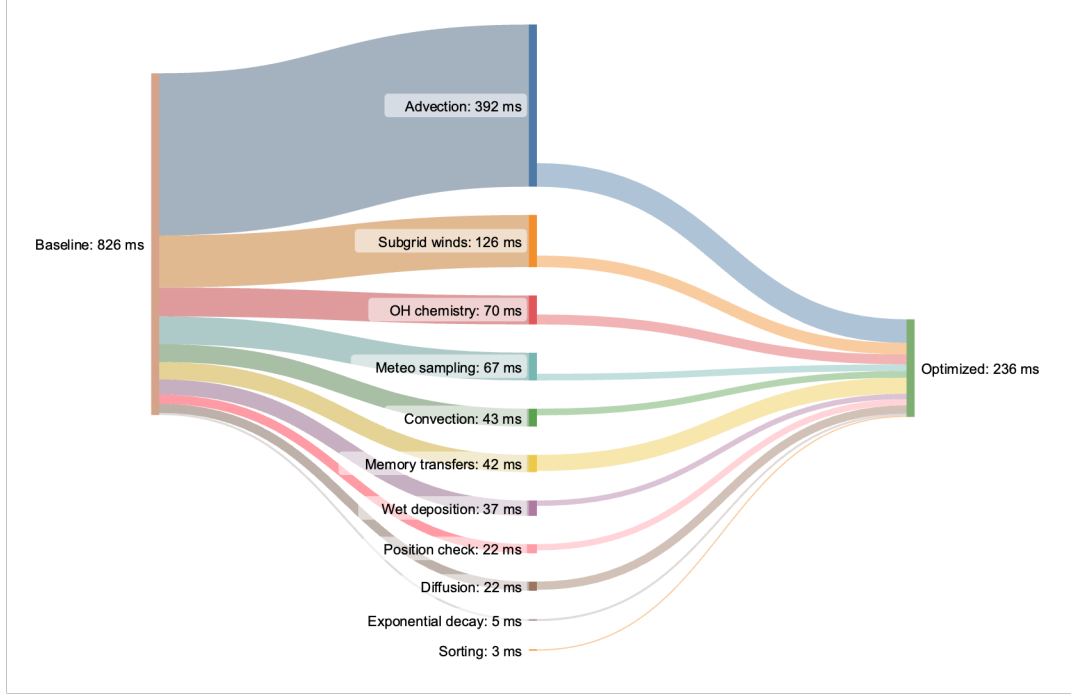


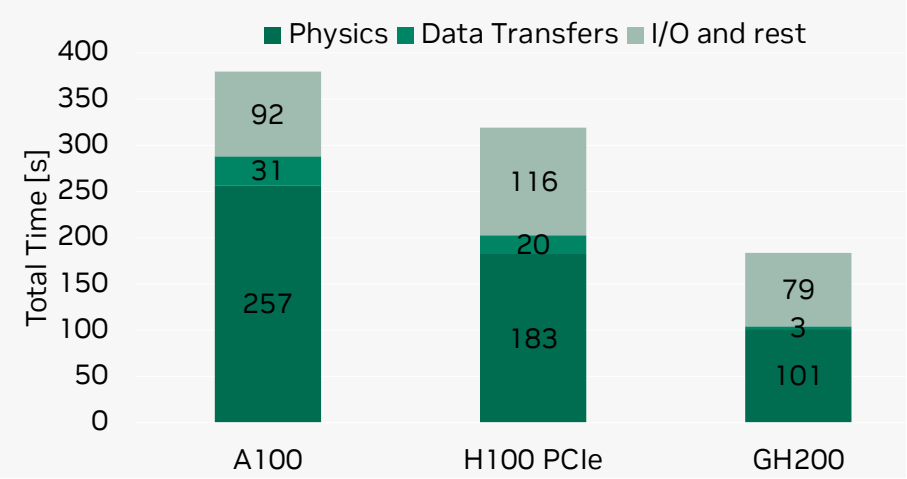
https://github.com/OpenGPTX/Megatron-LM/

MPTRAC

By Lars Hoffmann / JSC

- Lagrangian particle dispersion model: atmospheric transport processes (troposphere/stratosphere)
  → volcanic emissions
- Continuously optimized for GPUs
  Recently: Significant speedup on A100
- First test on GH200



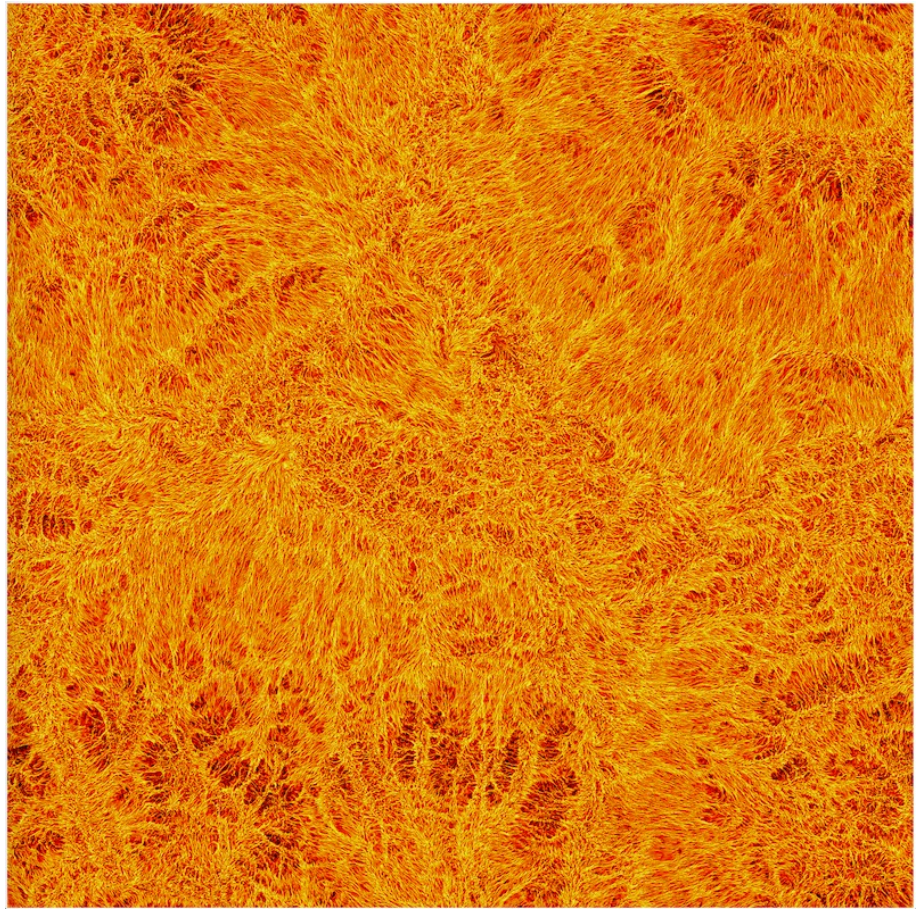


See also GTC talk by Mathias Wagner

OTHER GH200 RESULTS

JUWELS Booster A100 vs. GH200

- ICON:
  - R2B4 Benchmark: 563 s vs. 343 s → 1.62×
  - By Manoel Römmer / JSC
- nekRS
  - Rayleigh–Bénard Convection: 2.16×
  - By Mathis Bode / JSC
- Arbor
  - Busyring Benchmark: 330 s vs 167 s → 1.97×
  - By Thorsten Hater / JSC

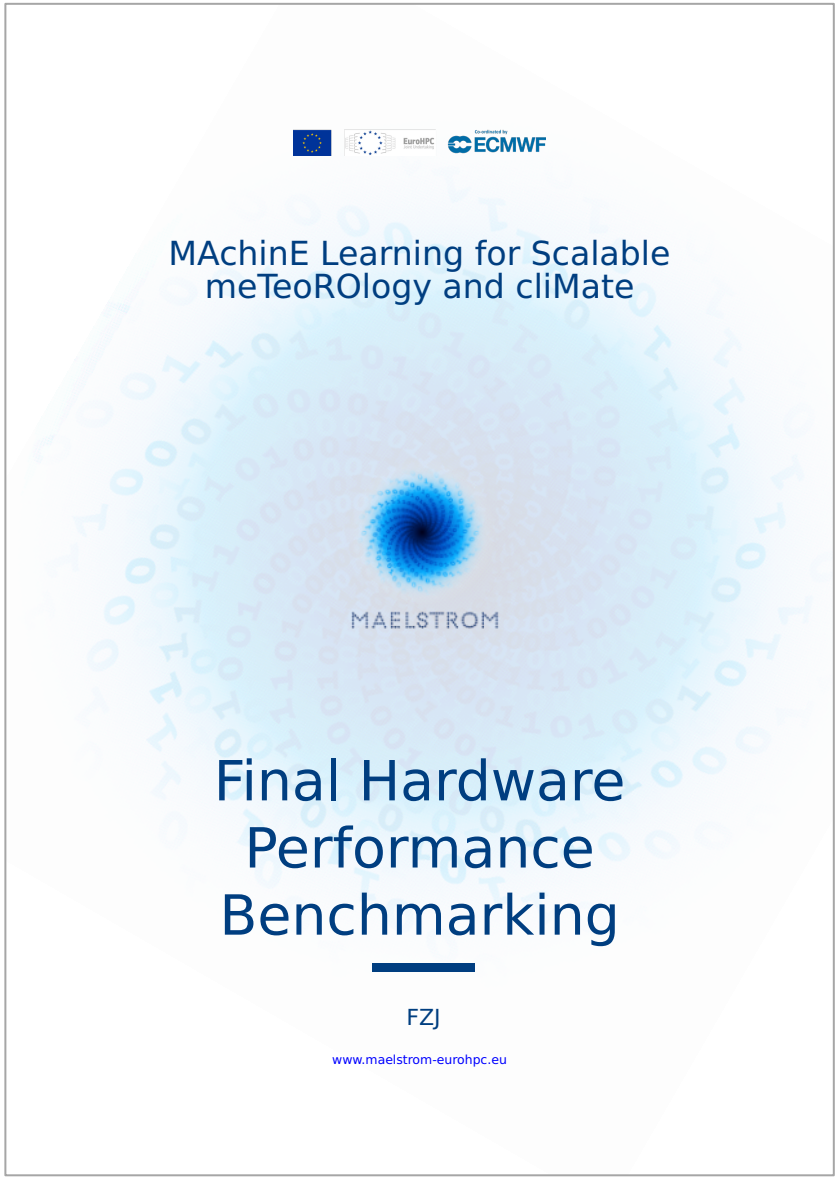


RBC Visualization (M. Bode)

MAELSTROM TESTS

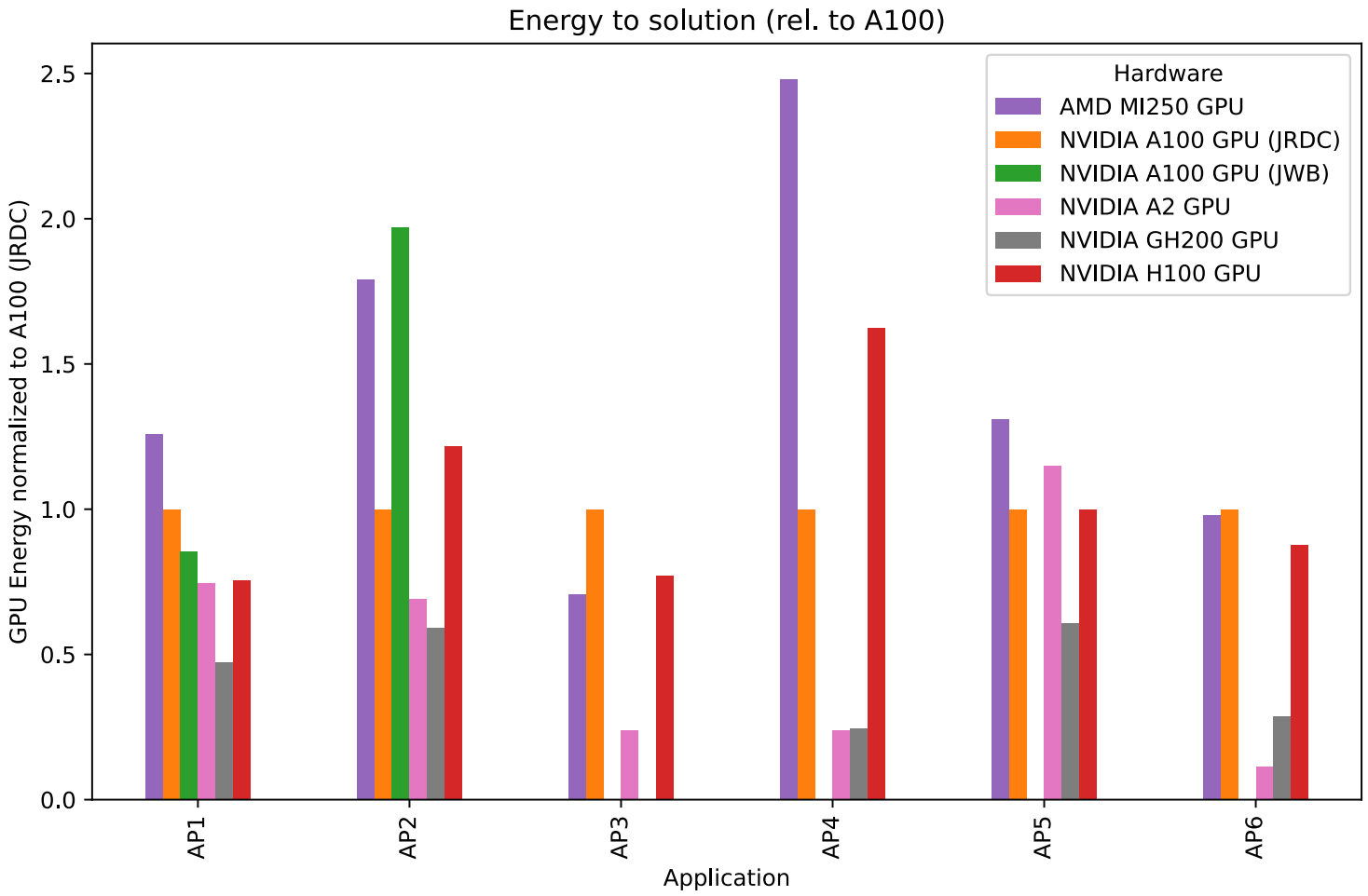
- MAELSTROM: EuroHPC-JU-funded project, ended in March 2024
- Enablement/optimization of 6 weather & climate applications using ML
- D3.7 (Feb 2024): Benchmarking of final applications on diverse hardware
  - GH200 from E4, Italy

→ www.maelstrom-eurohpc.eu



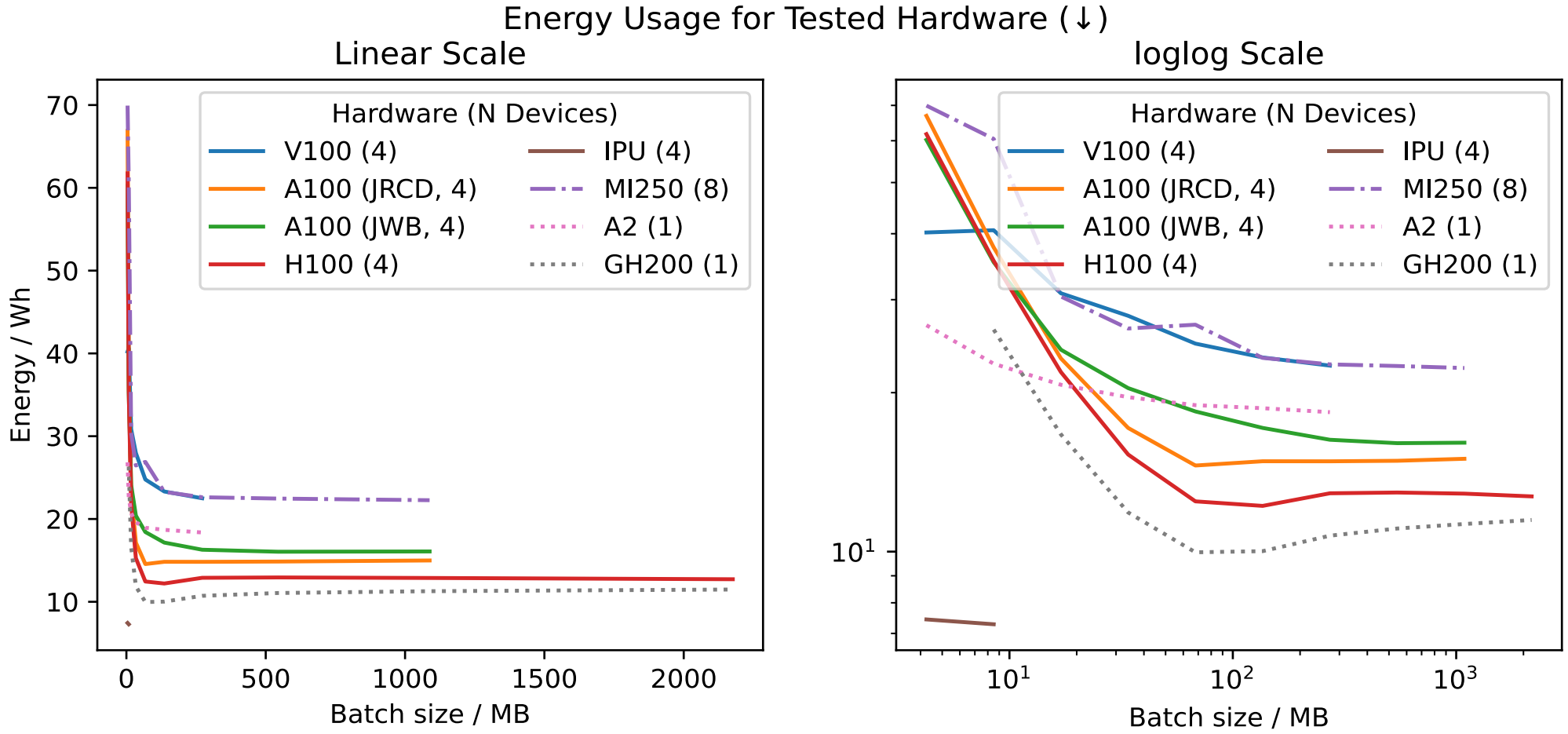
MAELSTROM TESTS

Energy-to-Solution Application Overview



MAELSTROM TESTS

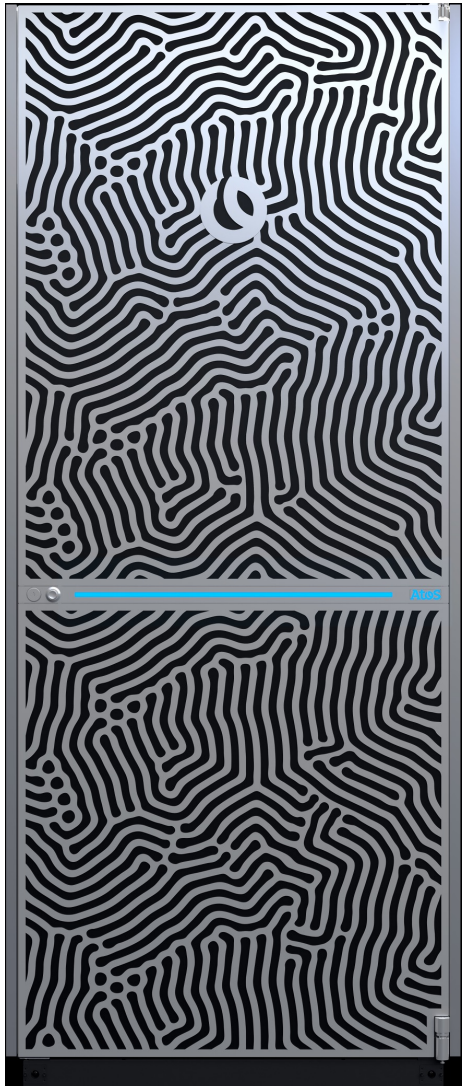
AP1: MetNor



MISSION BRIEFING OVERVIEW

- En route to JUPITER: EuroHPC JU system hosted at JSC
- Launched with focus on applications
- ~6000 nodes,
  24 000 H100 GPUs, 1 728 000 Arm cores, 24 000 NDR200 endpoints
- Landing in Modular Data Center
- Preparing for descent:
  - JUREAP
  - GH200 test systems

→ jupiter.fz-juelich.de



JUPITER

The Arrival of Exascale in Europe

fz-juelich.de/jupiter | #exa_jupiter

EuroHPC Joint Undertaking

Federal Ministry of Education and Research

Ministry of Culture and Science of the State of North Rhine-Westphalia

GCS Gauss Centre for Supercomputing