PAUL SCHERRER INSTITUT

Derek Feichtinger ::

HPCE Group presentation

SCD/AWI meeting, 7. Febuary 2023

PAUL SCHERRER INSTITUT PSI

Outline

PAUL SCHERRER INSTITUT PSI

Topic

PAUL SCHERRER INSTITUT

HPCE Group Members

Derek Feichtinger
Group Head
LHC Computing, Merlin6,
CSCS Ressources

Spencer Bliven
Systems Engineer
BIO Computing, Merlin6,
BIO Projects

Marc Caubet
Systems Engineer
Merlin6, MeG, CSCS
Ressources, Puppet

Elsa Germann
Systems Engineer / DevOps
TransAlps Project

Achim Gsell
Systems Engineer
SW Provis. (Pmodules),
AFS, Projects, ...

Hans-Nikolai Viessmann
Systems Engineer / DevOps
TransAlps Project

HPC Resources offered by HPCE - PSI hosted systems

Providing and offering access for PSI users to various HPC / HTC systems and services

- Merlin6: open to all PSI users
  - some resources owned by specific stakeholders.

HPC Resources offered by HPCE - PSI hosted systems

Providing and offering access for PSI users to various HPC / HTC systems and services

- Merlin6: open to all PSI users
  - some resources owned by specific stakeholders.
- MeG: Dedicated resources for MeG Experiment Collaboration (NUM)
  - offline cluster for MeG, shares Merlin storage

PAUL SCHERRER INSTITUT

HPC Resources offered by HPCE - PSI hosted systems

Providing and offering access for PSI users to various HPC / HTC systems and services

- Merlin6: open to all PSI users
  - some resources owned by specific stakeholders.
- MeG: Dedicated resources for MeG Experiment Collaboration (NUM)
  - offline cluster for MeG, shares Merlin storage
- LHC/CMS Tier-3: CMS experiment members of ETHZ, PSI, UniZ
  - connected to LHC Grid sites
  - system located in DMZ

HPC Resources offered by HPCE - PSI hosted systems

Providing and offering access for PSI users to various HPC / HTC systems and services

- Merlin6: open to all PSI users
  - some resources owned by specific stakeholders.
- MeG: Dedicated resources for MeG Experiment Collaboration (NUM)
  - offline cluster for MeG, shares Merlin storage
- LHC/CMS Tier-3: CMS experiment members of ETHZ, PSI, UniZ
  - connected to LHC Grid sites
  - system located in DMZ

All these systems are planned to get implemented on top of CSCS Alps *vClusters* in the future. (TransAlps project)

Merlin6 overview

- 4 HPE Apollo k6000 enclosures
- Each enclosure hosting 24 HPE ProLiant XL230k Gen10 Servers
- EDR Infiniband 100 Gb/s Switch system

k6000 enclosure

Merlin6 overview

- 4 HPE Apollo k6000 enclosures
- Each enclosure hosting 24 HPE ProLiant XL230k Gen10 Servers
- EDR Infiniband 100 Gb/s Switch system
- Each server with 2 Xeon Gold Skylake CPUs
  - 2 * Xeon Gold 6152 (44 cores) / 6240R (48 cores):
  - 384/768 GB RAM
  - 1.6 TB NVMe local disk
- Total cores of Merlin6 CPU multicore nodes: 4384 cores

k6000 enclosure

Merlin6 overview

- 4 HPE Apollo k6000 enclosures
- Each enclosure hosting 24 HPE ProLiant XL230k Gen10 Servers
- EDR Infiniband 100 Gb/s Switch system
- Each server with 2 Xeon Gold Skylake CPUs
  - 2 * Xeon Gold 6152 (44 cores) / 6240R (48 cores):
  - 384/768 GB RAM
  - 1.6 TB NVMe local disk
- Total cores of Merlin6 CPU multicore nodes: 4384 cores
- 15 nodes with consumer grade GPUs (62 GPUs) procured by BIO
  - Tesla K80, GTX 1080, GTX 1080 Ti, RTX 2080 Ti
- 1 high end NVidia DGX-A100 node "Gwendolen" with 8 A100 GPUs connected by NVLink

k6000 enclosure

PAUL SCHERRER INSTITUT

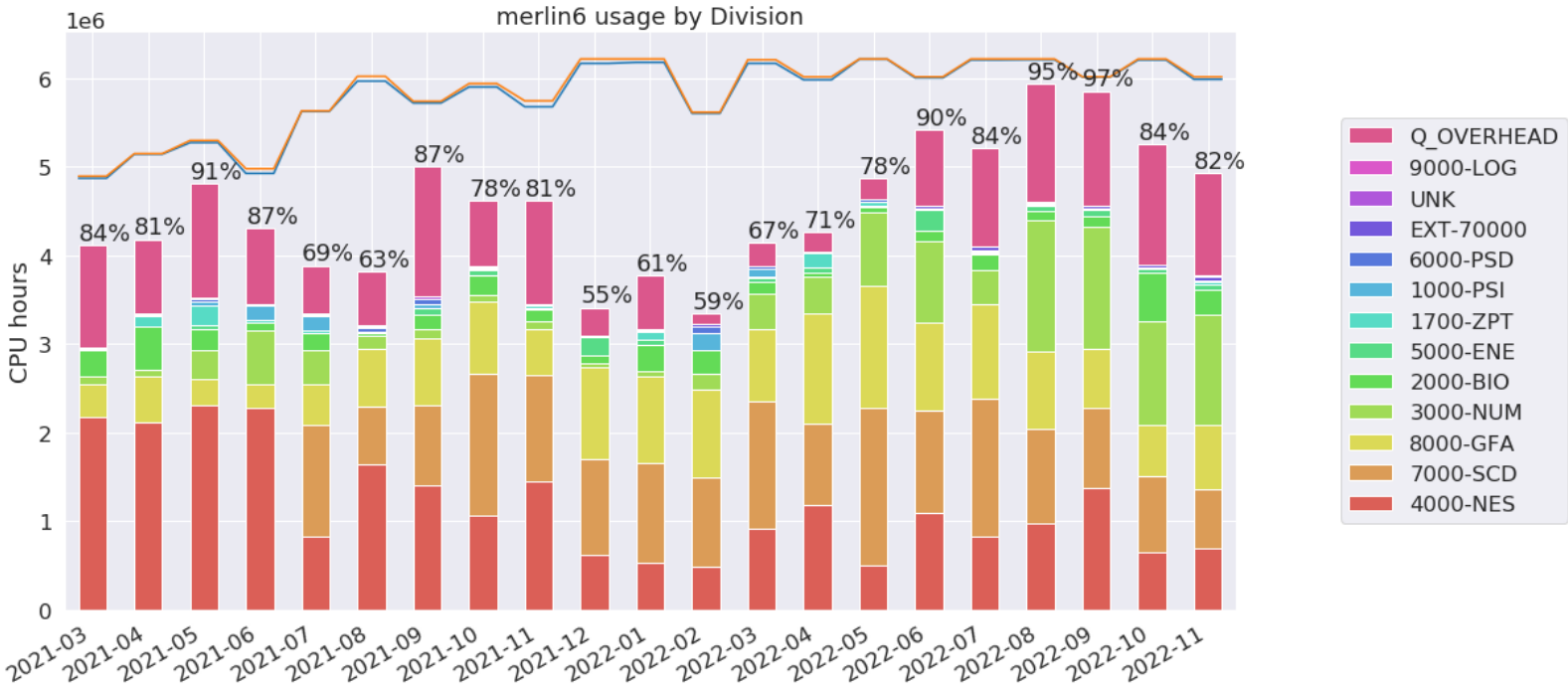
Merlin6 Investments by Division

Source	kCHF	%	Components
Director	1560.68	57.5	66 nodes, Storage
GFA	555.74	20.5	23 nodes
NUM Mu3e	185.93	6.8	6 nodes, Storage
NUM MEG	109.00	4.0	Storage
BIO	134.86	5.0	GPU nodes
BeAufw Merlin	92.43	3.4	Operations
LSM	47.41	1.7	1 node, DGX-A100
ZPT	20.00	0.7	DGX-A100
NES	10.00	0.4	DGX-A100
TOTAL	2716.05	100.0	

- Cluster lifetime: usually 5 years, bounded by storage
  - Storage HW usually shows strong decline after 5 years. Buying with 5 years warranty is customary option.
  - Compute nodes often run beyond warranty

PAUL SCHERRER INSTITUT

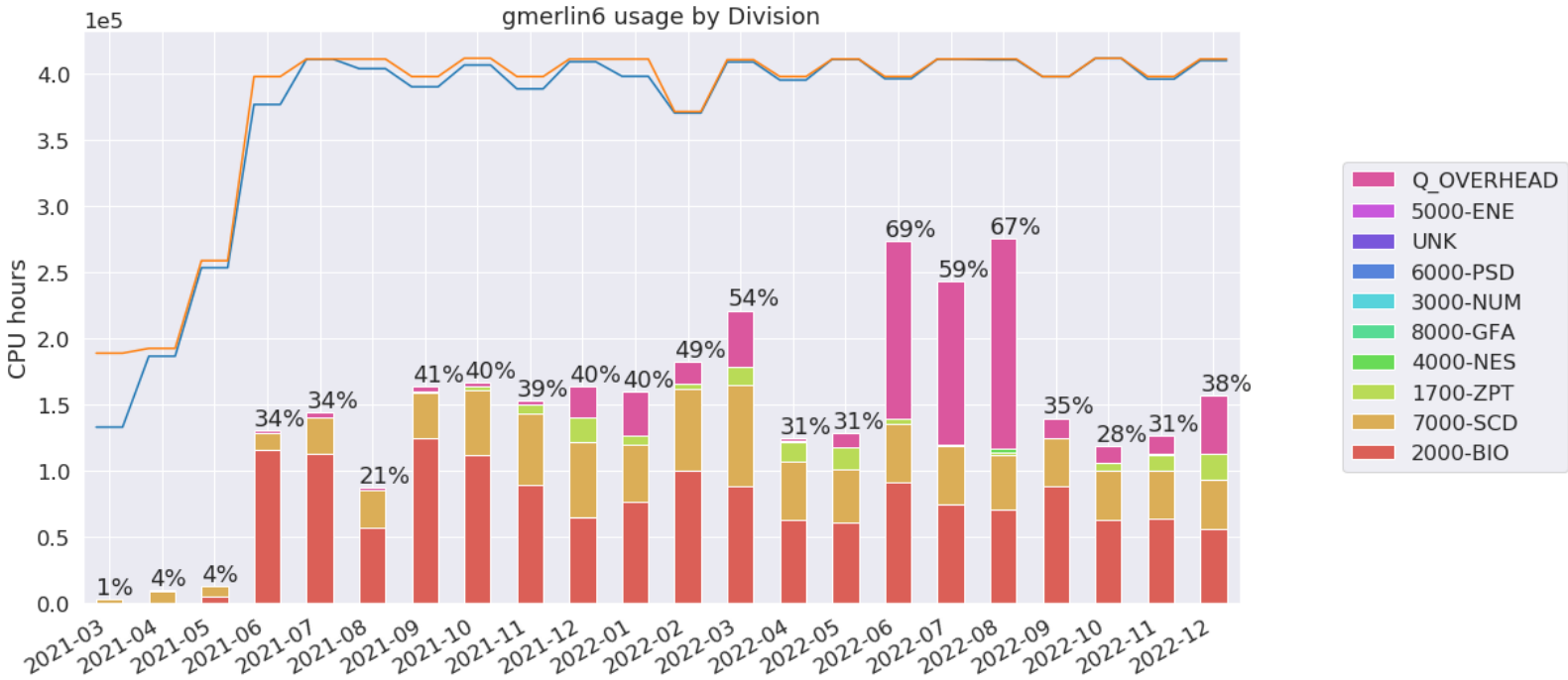
Merlin Usage Statistics per Division



- Usage calculated in respect to all available cores of the system (mixed workloads per node)
- the top lines indicate the total amount of resources, and how many have been available per month.

PAUL SCHERRER INSTITUT

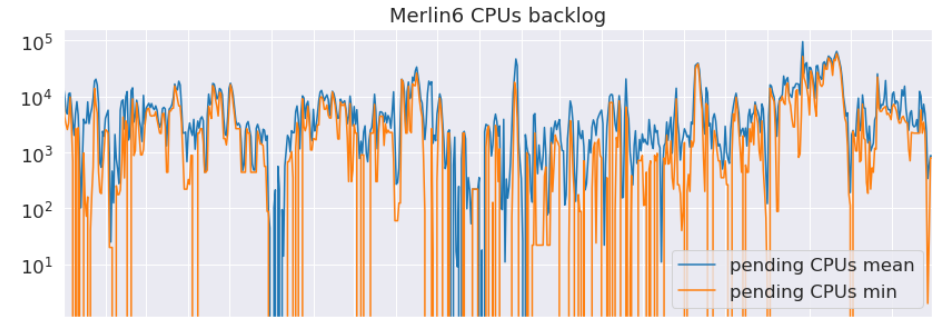
Merlin GPU node Usage per Division

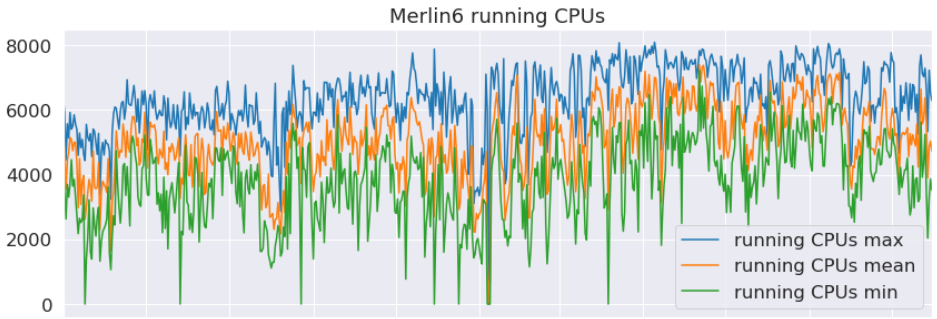


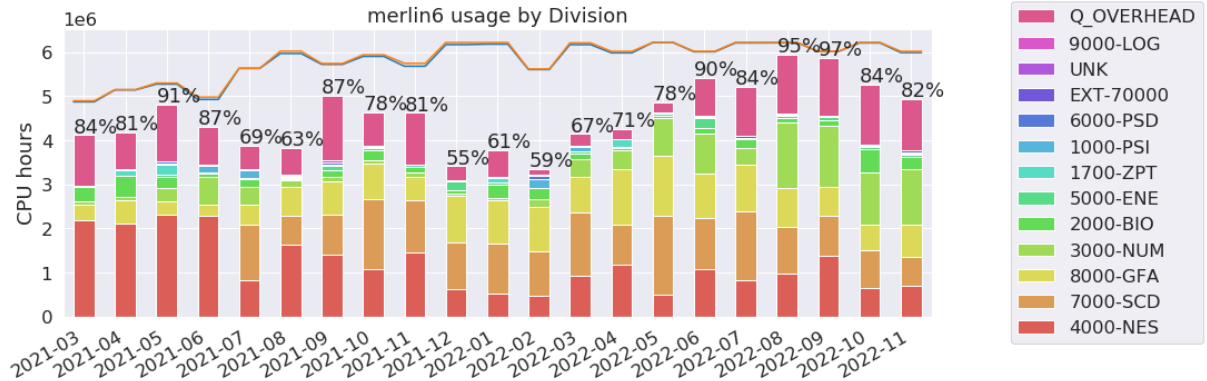
- NOTE: This only counts the CPU usage and not the GPU usage/reservation
- can only be taken as an approximation, still I want to show it to give an approximate impression about the usage by division
- includes consumer grade GPU nodes + the restricted Gwendolen DGX-A100
- consumer grade GPU nodes usually fully occupied

PAUL SCHERRER INSTITUT

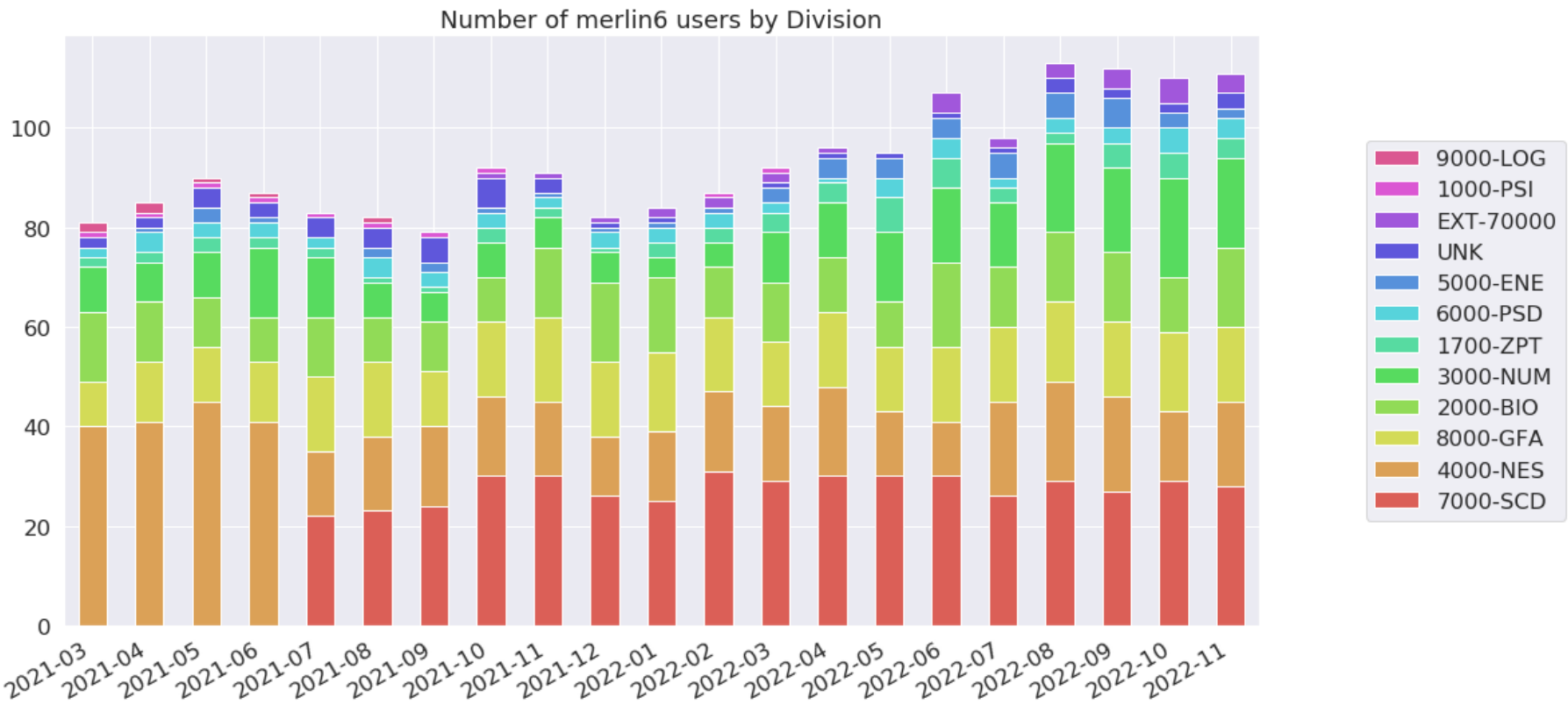
Merlin6 usage and backlog







Active Merlin Users per Division



262 users, 85 research groups and all 9 PSI divisions (+ external guests) over the time range

HPC Resources offered through HPCE - external systems

- CSCS Piz Daint: open to all users
  - requires a yearly project application. Projects with a potential of entering CSCS competitive application should use the competitive process.

PAUL SCHERRER INSTITUT

HPC Resources offered through HPCE - external systems

- CSCS Piz Daint: open to all users
  - requires a yearly project application. Projects with a potential of entering CSCS competitive application should use the competitive process.
- LHC CSCS Tier-2: currently on Piz Daint / Alps
  - Run by CSCS for CHIPP community. I am acting as site contact for CMS to help operate and define the services

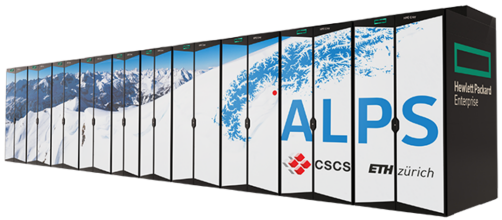
PAUL SCHERRER INSTITUT PSI

HPC Resources offered through HPCE - external systems

- CSCS Piz Daint: open to all users
  - requires a yearly project application. Projects with a potential of entering CSCS competitive application should use the competitive process.
- LHC CSCS Tier-2: currently on Piz Daint / Alps
  - Run by CSCS for CHIPP community. I am acting as site contact for CMS to help operate and define the services

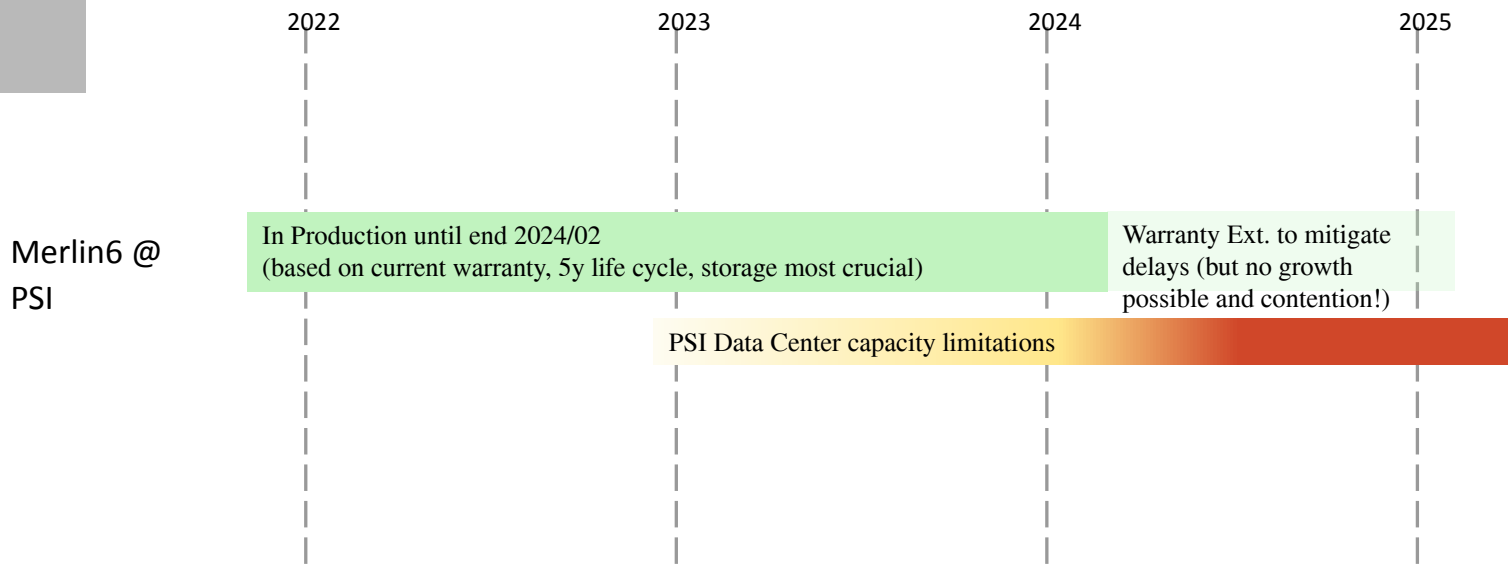
TransAlps project

Future HPC/HTC Resources at CSCS Alps. Merlin7 cluster as first implementation



TransAlps Timeline Update

Current PSI Merlin6 Production System Timeline

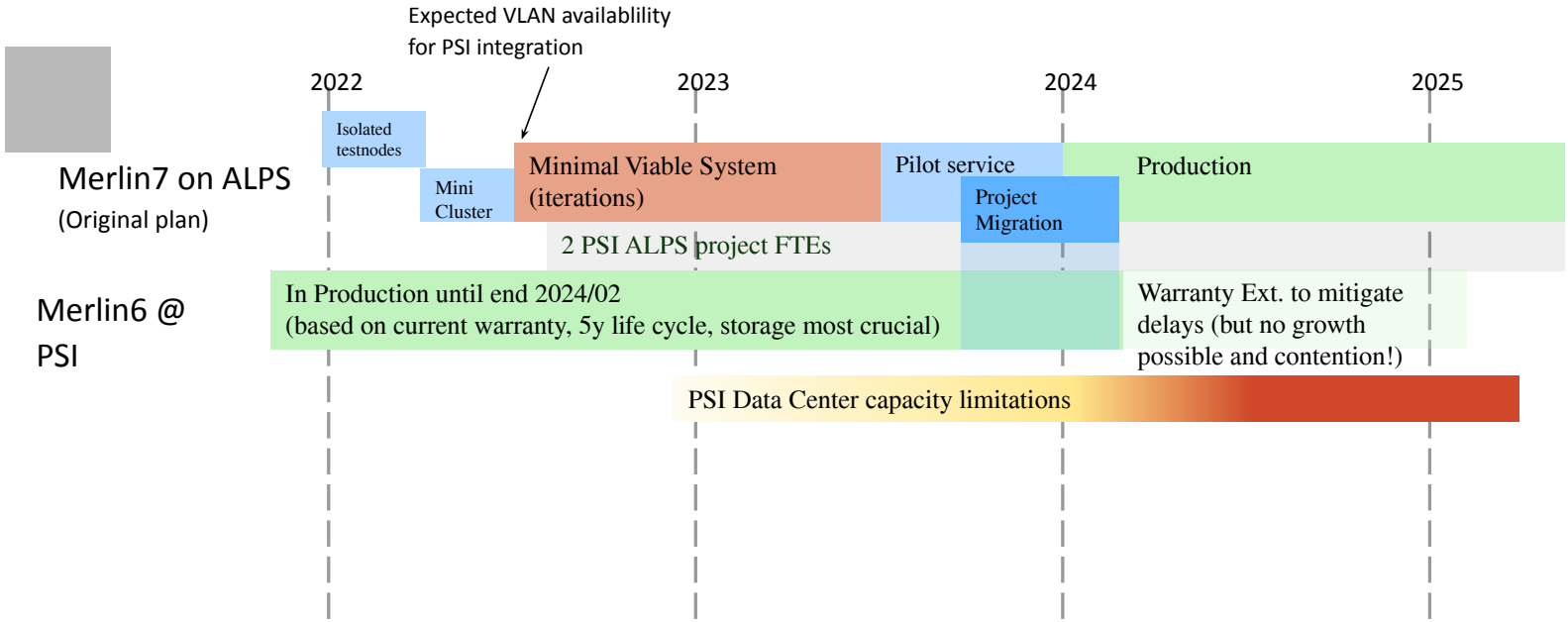


CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Page 7

TransAlps Timeline Update

TransAlps Project Timeline

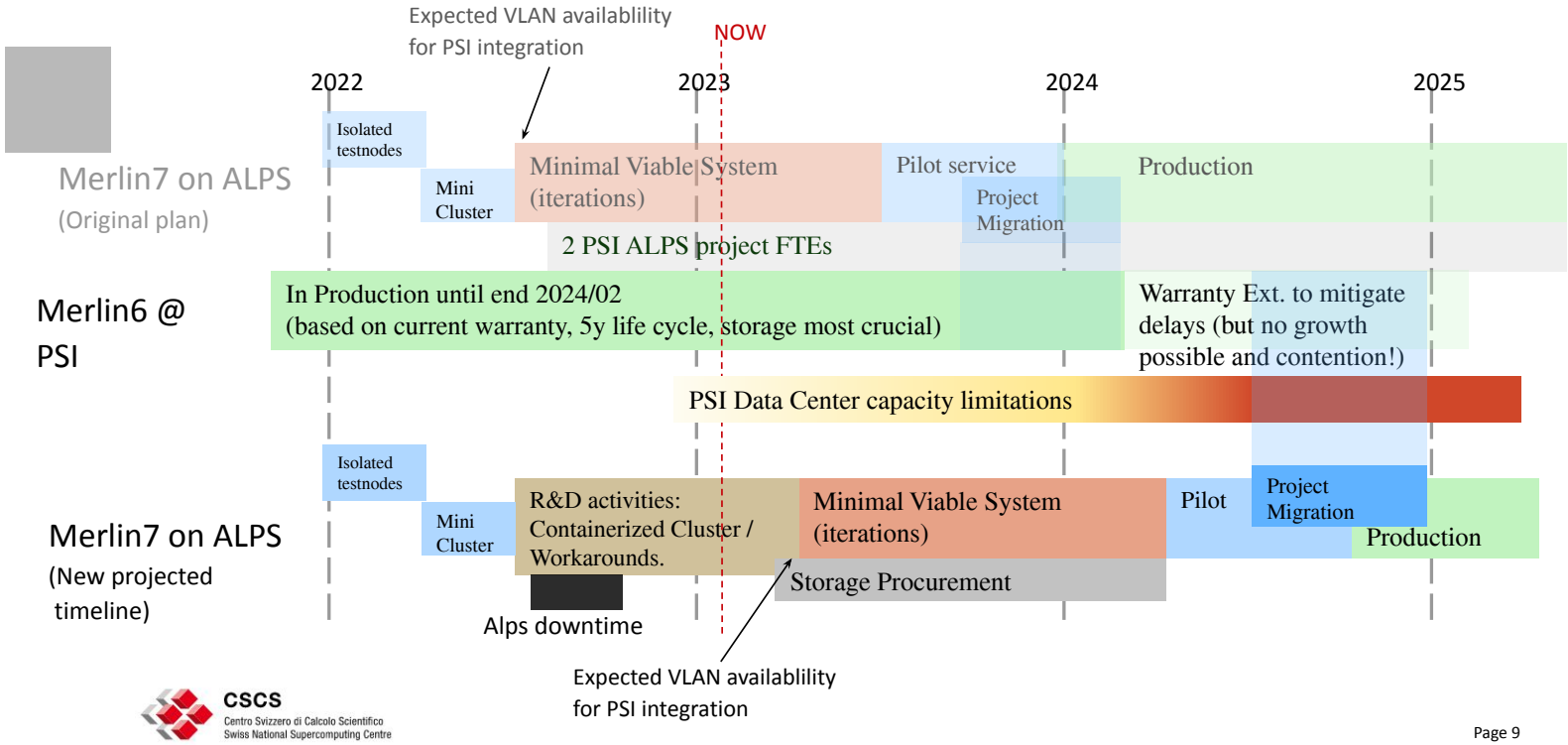


CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Page 8

TransAlps Timeline Update

TransAlps Project Timeline



CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Page 9

PAUL SCHERRER INSTITUT

Marc Caubet

- Main architect and system engineer for Merlin6
- Many essential contributions to central PSI Linux infrastructure
  - Puppet (especially HPC related modules)
  - Slurm scheduler
  - maintains many SW builds inside of the Pmodule system
- PSI Piz Daint resources - co-administrating
- Deputy for Tier-3 operations

Current main project

PAUL SCHERRER INSTITUT PSI

Topic

PAUL SCHERRER INSTITUT

Merlin6 Compute Hardware 2023

Description (Owner)	No. of nodes	Name merlin-*	Processors	GPU	Cores /node	Memory [GB]	Mem /Core [GB]	total Cores
login nodes	2	l-001..002	2 Xeon Gold 6152		32	512	16.0	64
computing nodes	72	c-001..224	2 Xeon Gold 6152		44	384	8.7	3168
computing nodes	18	c-301..018	2 Xeon Gold 6240R		48	768	16.0	864
computing nodes	6	c-319..024	2 Xeon Gold 6240R		48	384	8.0	288
merlin5 nodes	29	c-18..47	2 Xeon E5-2670		16	64	4.0	464
GPU node	1	g-40	Xeon E5-2690 v3	4 Tesla K80	24	512	21.3	24
GPU node	1	g-001	Xeon E5-2640 v4	2 GTX 1080 Ti	20	128	6.4	20
GPU node	4	g-002..005	Xeon E5-2640 v4	4 GTX 1080	20	128	6.4	80
GPU node	4	g-006..009	Xeon E5-2640 v4	4 GTX 1080 Ti	20	128	6.4	80
GPU node	4	g-010..013	Xeon Silver 4210R	4 RTX 2080 Ti	20	128	6.4	80
GPU node	1	g-014	Xeon Silver 6240R	8 RTX 2080 Ti	24	384	16.0	24
Gwendolen GPU	1	g-100	2 x AMD EPYC 7742	8 A-100	128	1000	7.8	128
	143							5284

Spectrum Scale (GPFS) based HPC storage

Storage Allocation	PB
BIO (centr. funded)	2.3
general PSI users	1.4
Mu3e	1.3
MEG	1.2
	6.2