

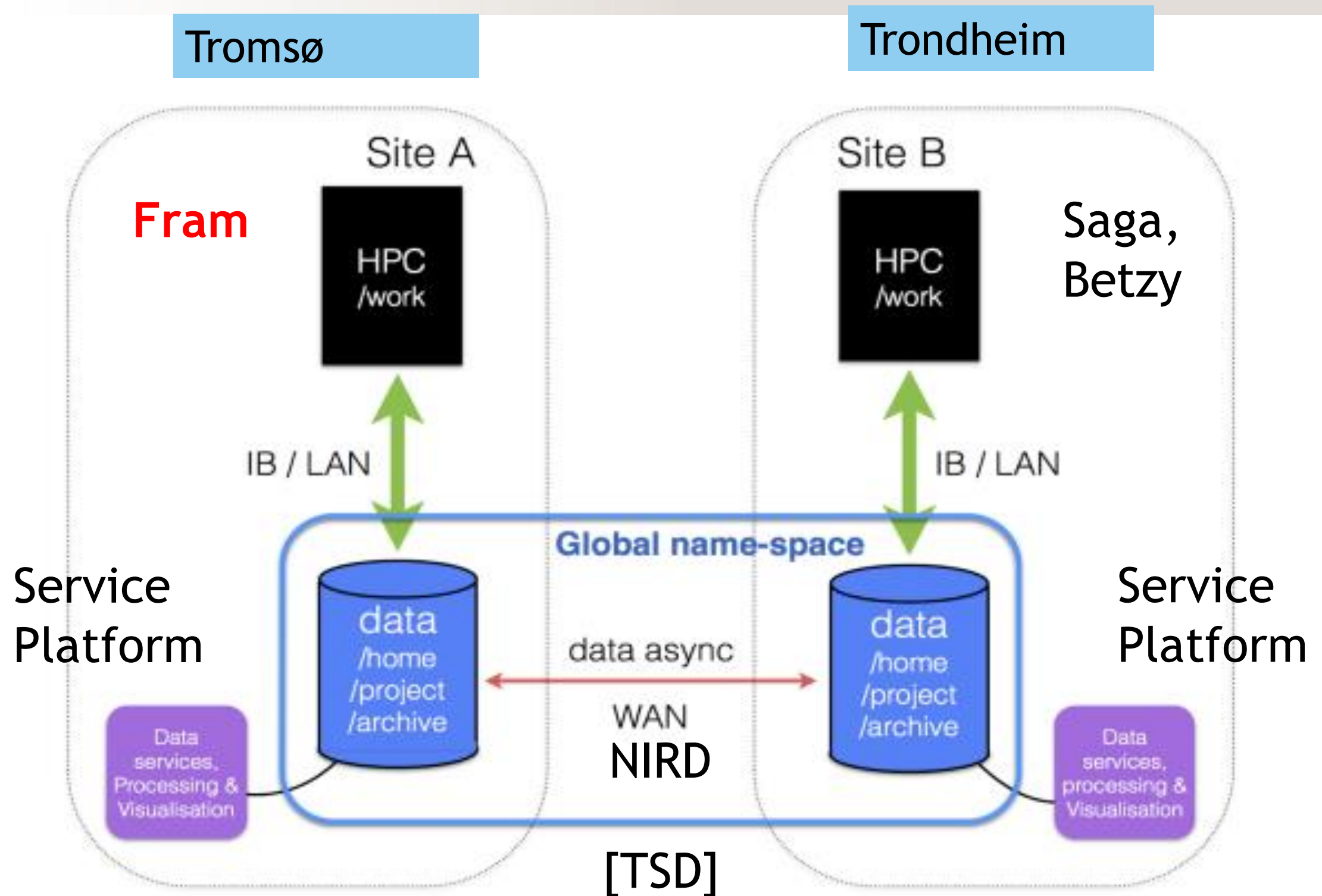
National e-infrastructure GPU services

Jørn Amundsen, UNINETT Sigma2 AS

2020-04-01



E-infrastructure implementation



HPC service GPU hardware

- Attached to the Saga system for bioinformatics and small scale I/O intensive workloads
- Providing GPU service for GPU enabled batch processing HPC applications through Slurm
- Hardware:
 - 200 standard compute nodes w/ 40 cores and 192 GiB RAM per node
 - 28 medium memory compute nodes w/40 cores and 384 GiB RAM/node
 - 8 big memory nodes w/64 cores and 3 TiB RAM/node
 - 8 GPU nodes with 2 CPUs (24 cores, 384 GiB RAM) and 4 NVIDIA P100 GPUs each

NIRD service platform GPU hardware

- ▶ Dockerized platform running Kubernetes
- ▶ Providing GPU service for AI/ML applications through its [deep-learning toolkit](#)
- ▶ Replicated service connected to the NIRD storage infrastructure
- ▶ Hardware:
 - Tromsø: 8 GPU nodes with 2 CPUs (32 cores, 256 GiB RAM) and 2 NVIDIA V100 GPUs each
 - Trondheim: 4 GPU nodes with 2 CPUs (40 cores, 768 GiB Ram) and 4 NVIDIA V100 GPUs each (also 4 CPU nodes w/o GPUs)

eX3 research platform

- ▶ Infrastructure to facilitate research on bleeding-edge HPC technologies
- ▶ Sigma2 is a partner in the consortium
- ▶ Featuring a DGX-2 system with 16 NVIDIA V100 GPUs
- ▶ Refer to ex3.simula.no for more information

LUMI GPU development platform

- It might be needed to provide more GPU resources while waiting for LUMI GPUS (main system or TDS) to become available
- Sigma2 will investigate options for providing intermediate resources for development
- Might need to look into what is the most probable GPU platform and cooperation with the GPU vendor
- As always open to community suggestions on development resources