MAX IV

Introducing Spectrum Scale at MAX IV Laboratory

- and lessons learned in the process

Andreas Mattsson

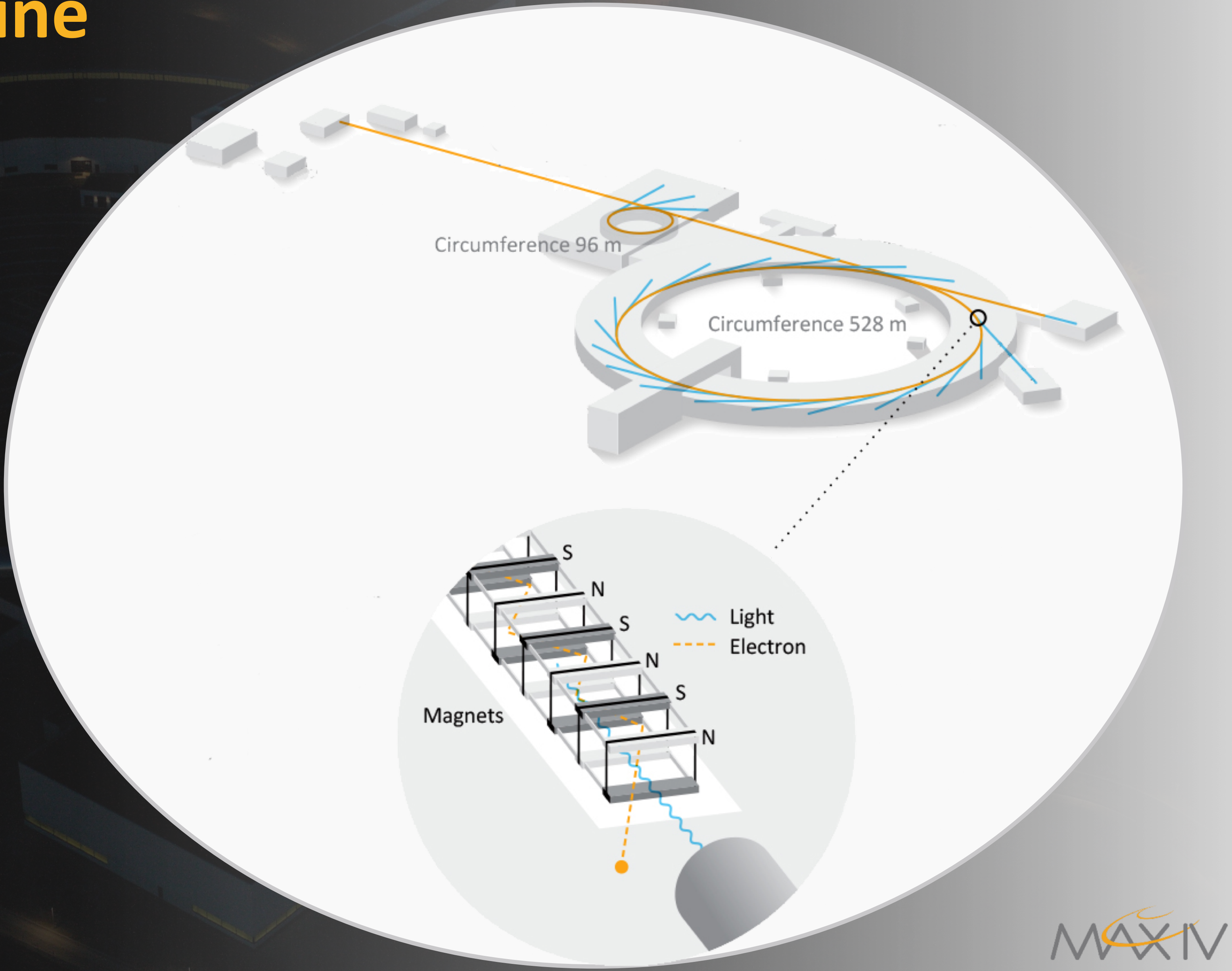
SSUG UK2018, London – 2018-04-19

What is MAX IV Laboratory?

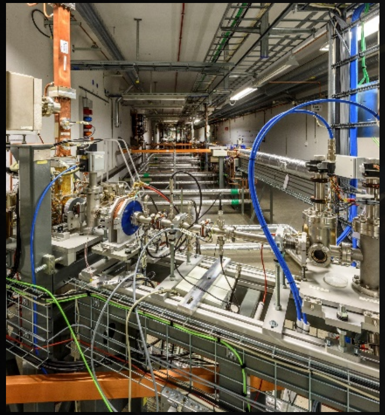
What is a Synchrotron?

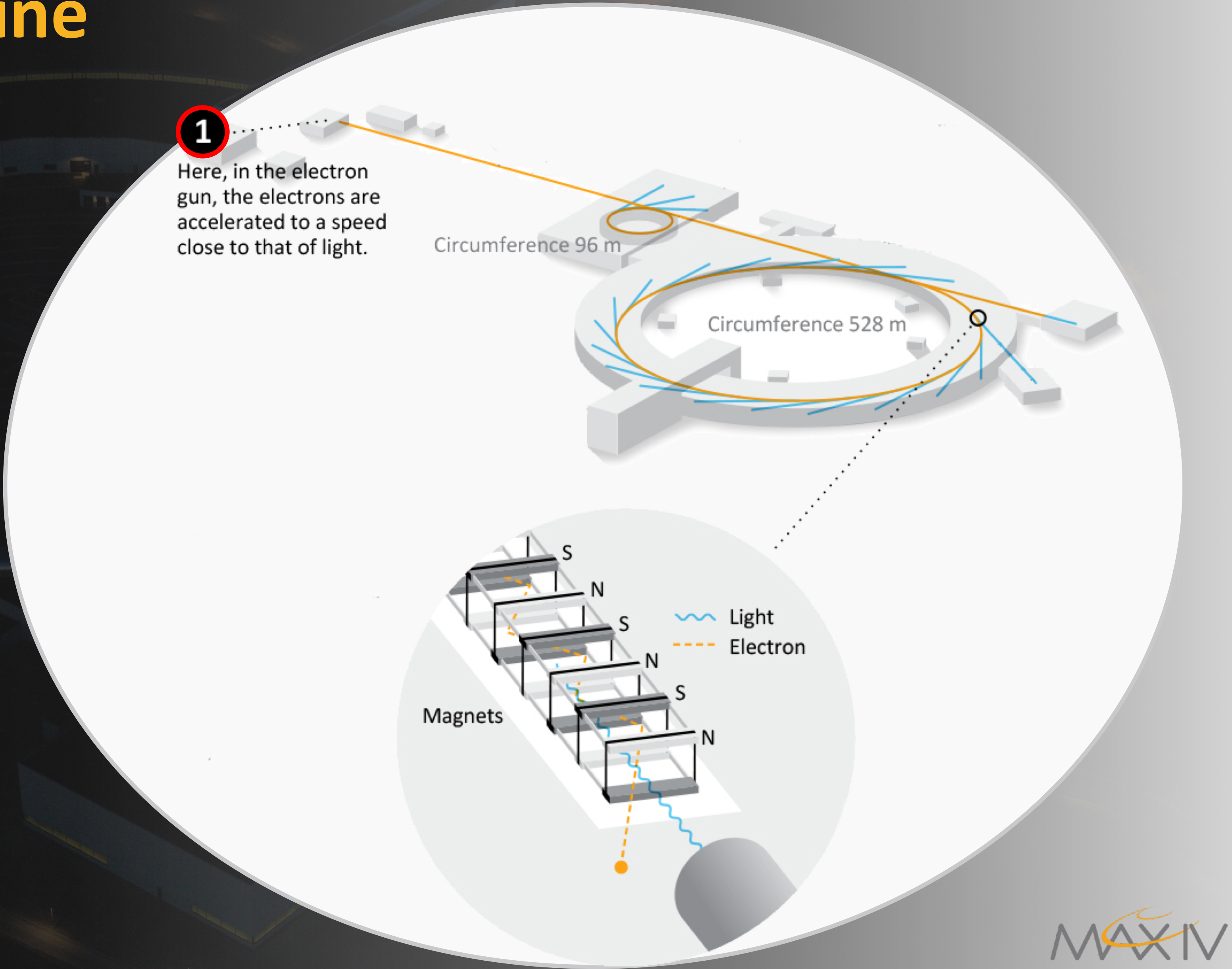
What is a Synchrotron?

The MAX IV Machine

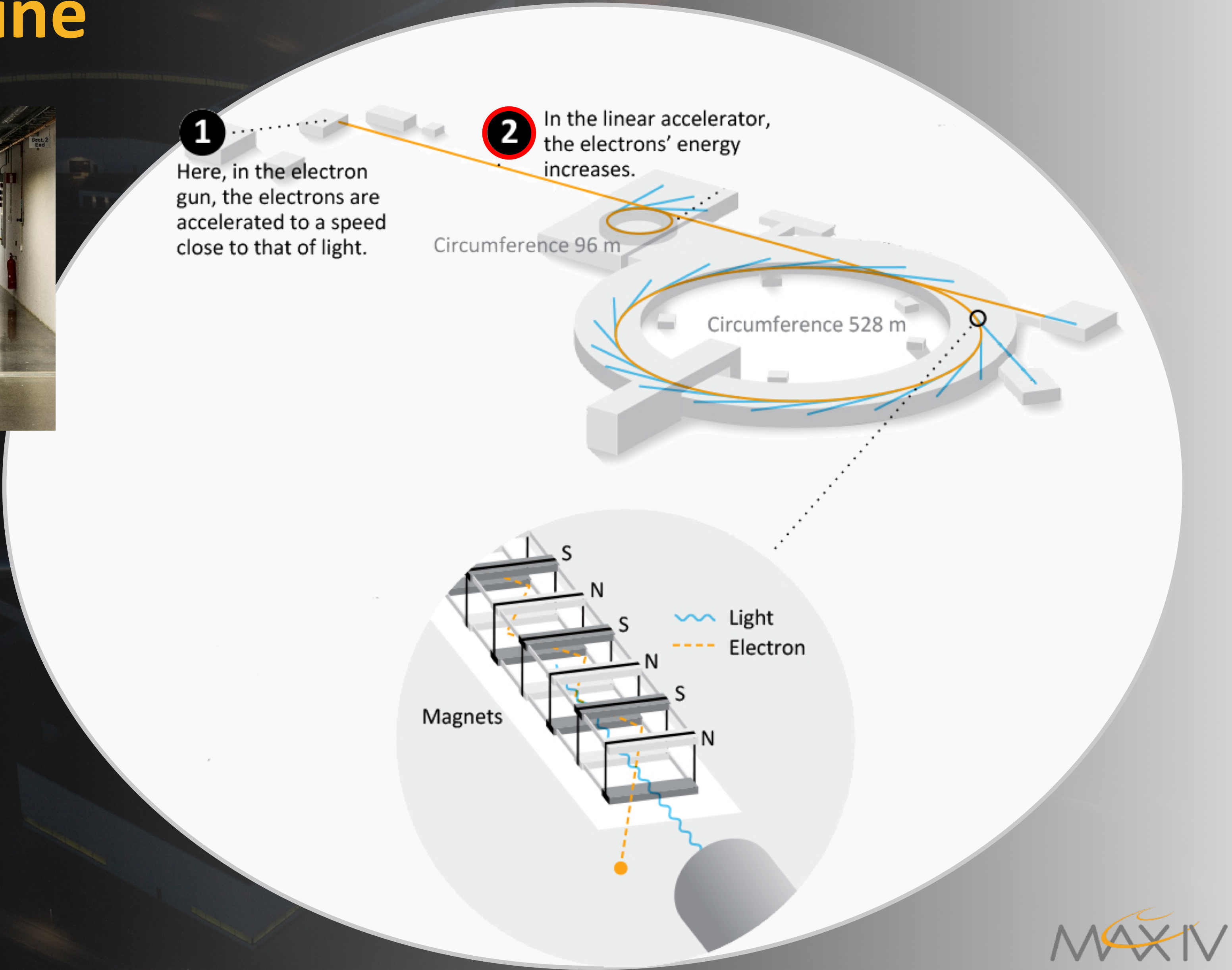


The MAX IV Machine




1
Here, in the electron gun, the electrons are accelerated to a speed close to that of light.
Circumference 96 m
Circumference 528 m
S
N
S
N
S
N
Light
Electron
Magnets

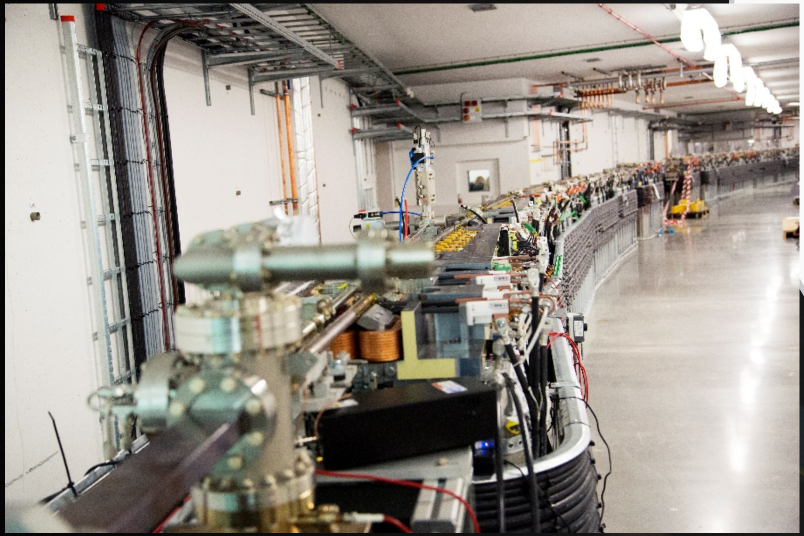
The MAX IV Machine

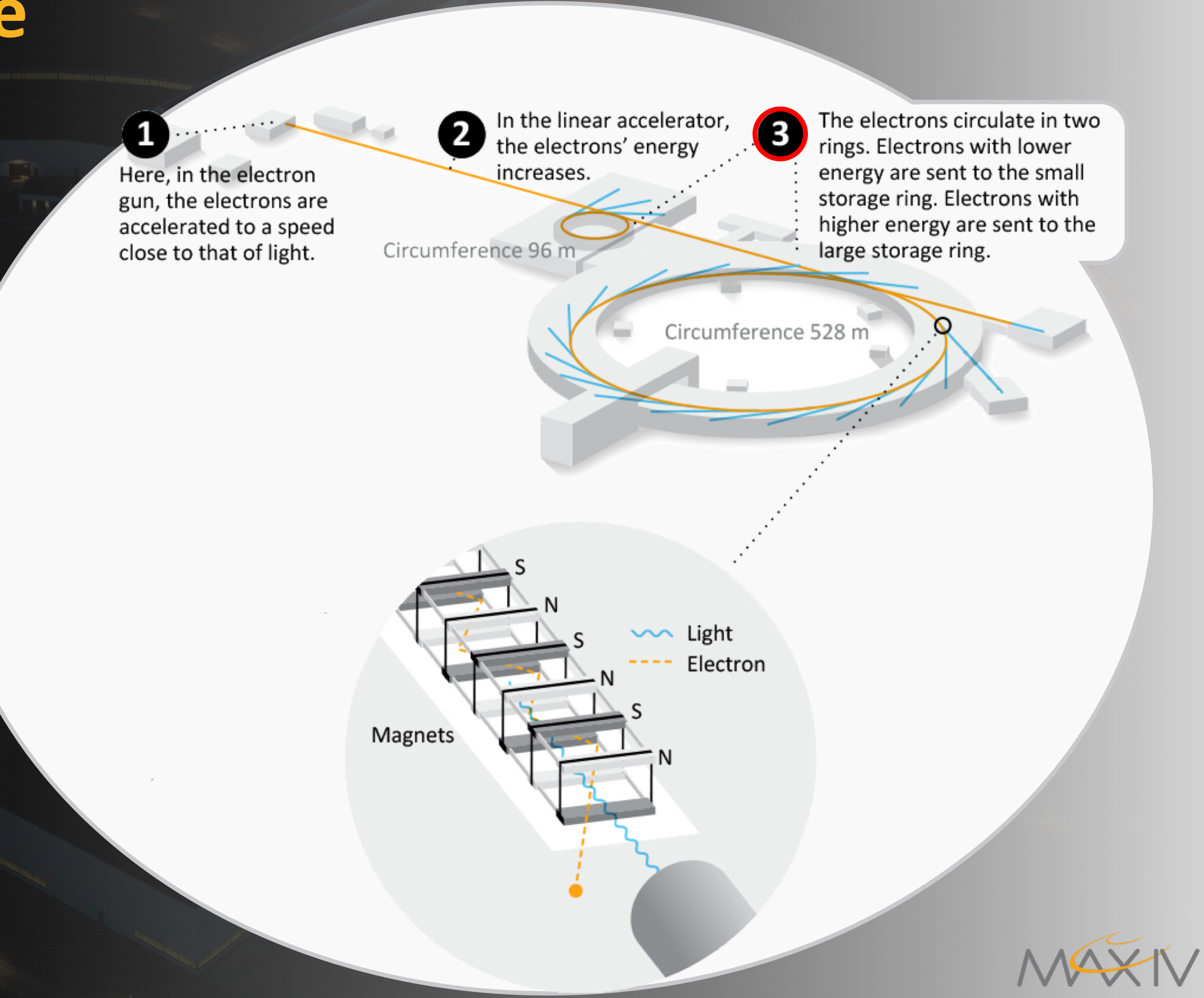


1 Here, in the electron gun, the electrons are accelerated to a speed close to that of light.

2 In the linear accelerator, the electrons' energy increases.

The MAX IV Machine



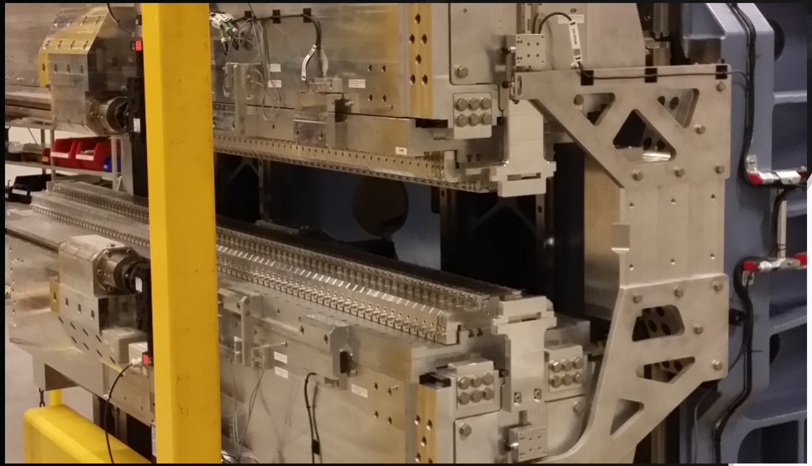


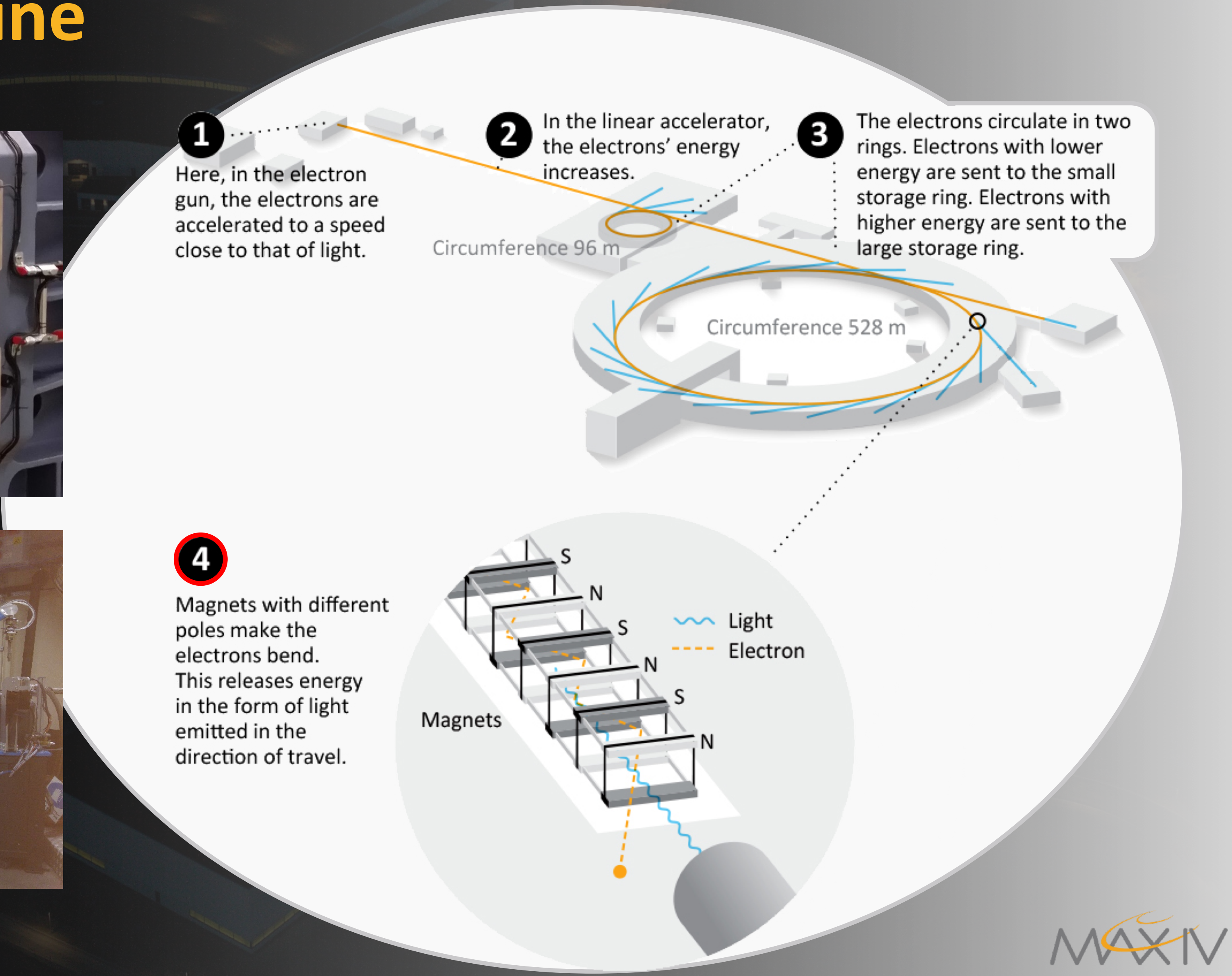
1 Here, in the electron gun, the electrons are accelerated to a speed close to that of light.

2 In the linear accelerator, the electrons' energy increases.

3 The electrons circulate in two rings. Electrons with lower energy are sent to the small storage ring. Electrons with higher energy are sent to the large storage ring.

The MAX IV Machine





1 Here, in the electron gun, the electrons are accelerated to a speed close to that of light.

2 In the linear accelerator, the electrons' energy increases.

3 The electrons circulate in two rings. Electrons with lower energy are sent to the small storage ring. Electrons with higher energy are sent to the large storage ring.

4 Magnets with different poles make the electrons bend. This releases energy in the form of light emitted in the direction of travel.

Why MAX IV Requires High Performance Storage

Current generation of detectors

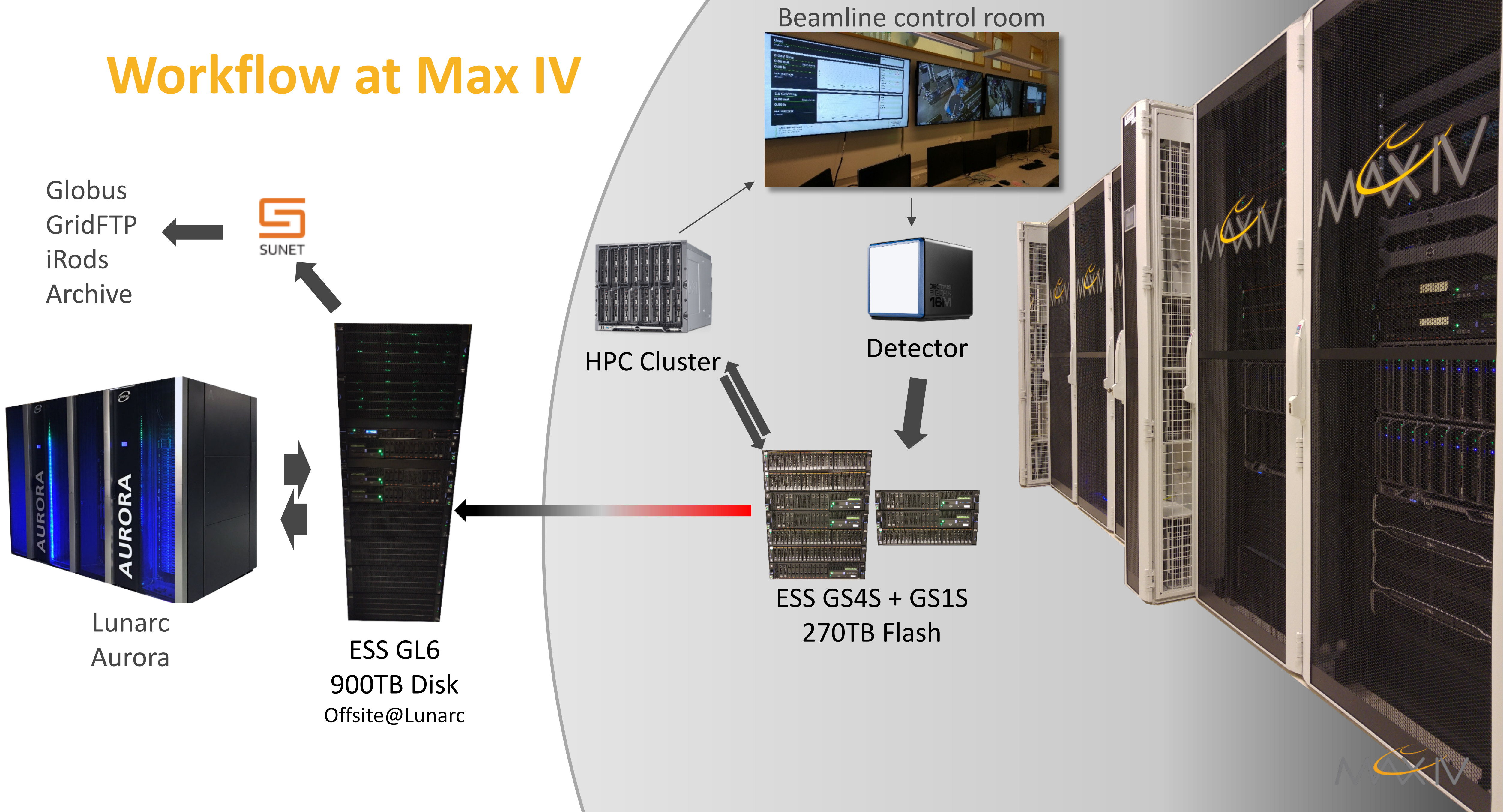
BioMAX

Detector:	Dectris Eiger X 16M
Resolution:	4150x4371x32 @ 133Hz
Native bitrate:	~ 70Gbps
Controller unit:	40Gbps

NanoMAX

Detector:	Dectris Eiger X 1M
Resolution:	1030x1065x32 @ 3000Hz
Native bitrate:	~100Gbps
Controller unit:	40Gbps

Workflow at Max IV



Globus
GridFTP
iRods
Archive

SUNET

Lunarc
Aurora

ESS GL6
900TB Disk
Offsite@Lunarc

Beamline control room

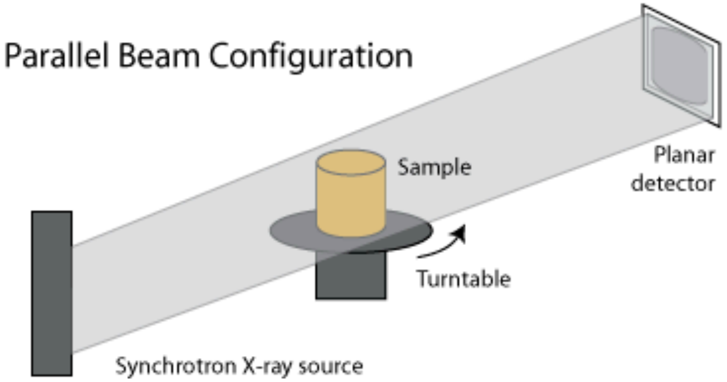
HPC Cluster

Detector

ESS GS4S + GS1S
270TB Flash

Tomographic Reconstruction

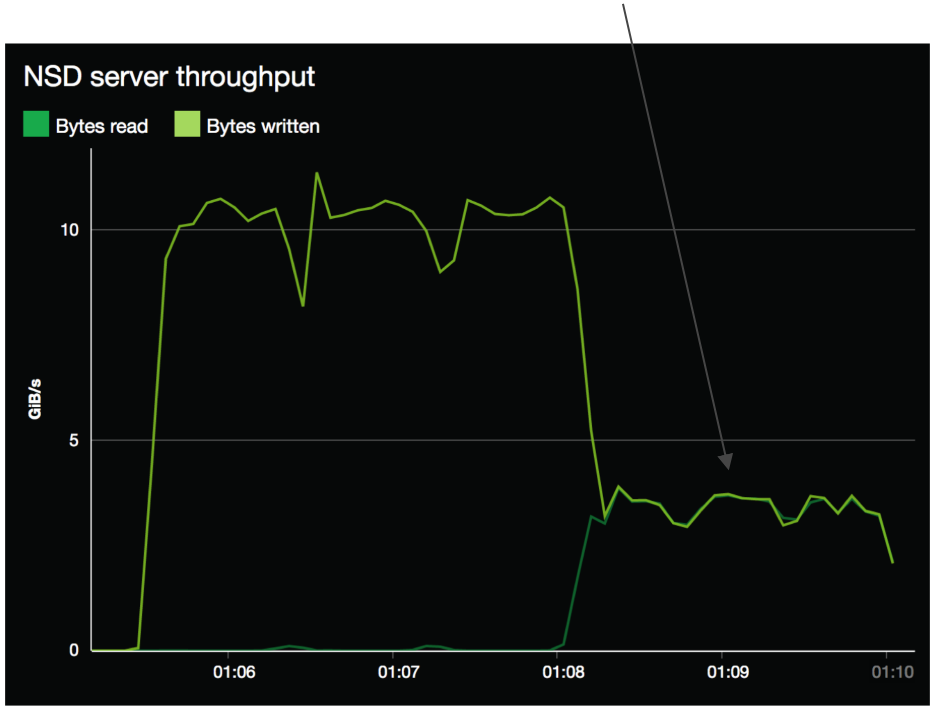
- Tomographic reconstruction is very well established method implemented in dedicated high performance algorithms and software



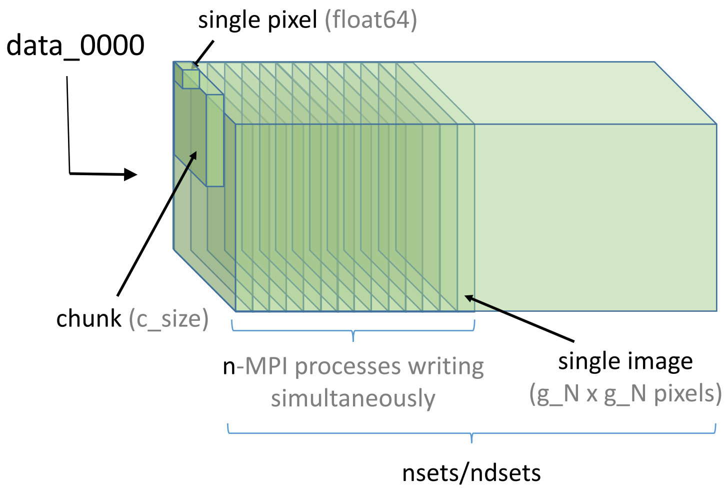
- Sample rotating, detector resolution 2048x2048, images for 1600 rotations and e.g. 1-100 time steps
  - single slice: 2048 x 2048 -> 16 Mbyte (single chunk of data)
  - Single measurement: x 1600 -> 25 GB
  - Time series: x 20-100 -> 0.5-2.5 TB (single file)
- HDF5 is MAX IV standard data format

Benchmark and Application cases

- **pwrite3dc: writing a time series of image like data into HDF5**
- with H5D_FILL_TIME_ALLOC (default settings for most of sw) there is simultaneous read/write affecting performance



Writing images to hdf5
Simply organized in a single dataset



data_0000

single pixel (float64)

chunk (c_size)

n-MPI processes writing simultaneously

single image (g_N x g_N pixels)

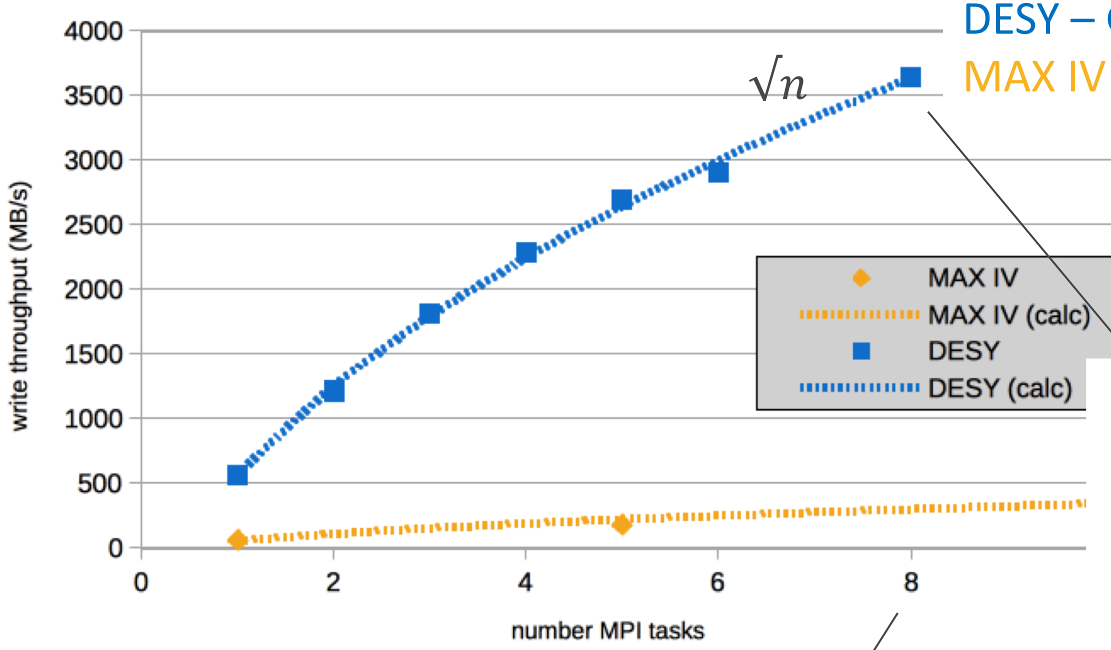
nsets/ndsets

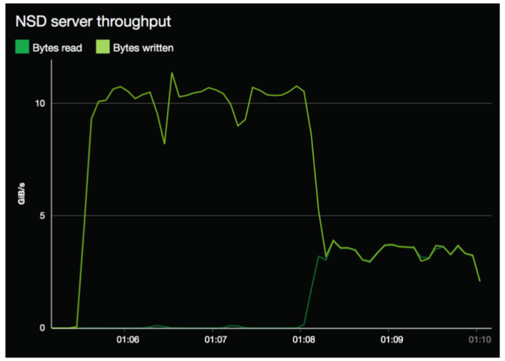
Comparison GL6 and GF1 with HDF5 – pwrite3dc

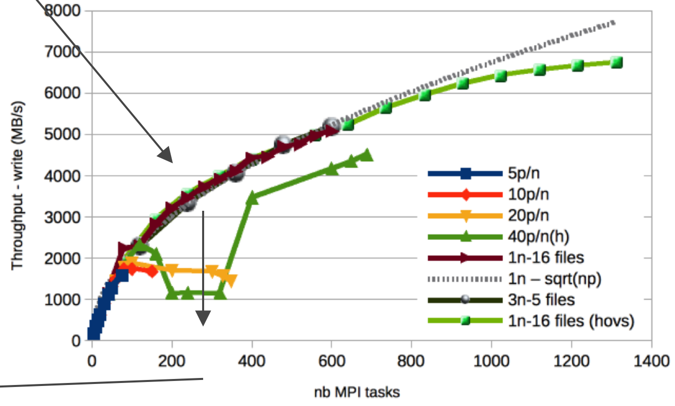
- gpfs, fill on alloc, cb-disabled, ds-enabled, single node test

DESY – GF1

MAX IV – GL6





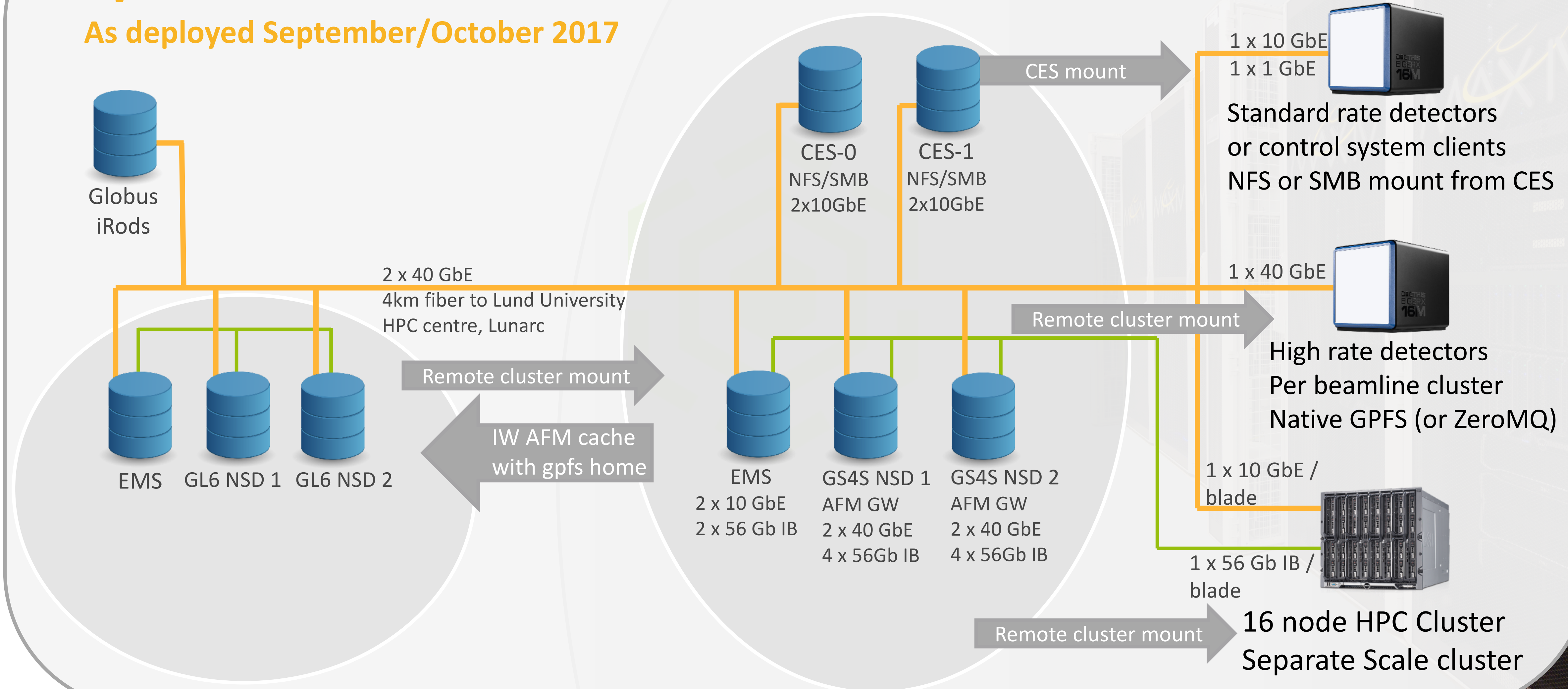


- with flash based GF1 (DESY) 12x faster
- with flash based GF1 (DESY) 40x less resources needed
  (4x hw., 10x sw.) to stretch the system for the same performance

Spectrum Scale infrastructure at MAX IV

Spectrum Scale infrastructure at MAX IV

As deployed September/October 2017



Globus
iRods

2 x 40 GbE
4km fiber to Lund University
HPC centre, Lunarc

EMS

GL6 NSD 1

GL6 NSD 2

Remote cluster mount

IW AFM cache
with gpfs home

CES-0
NFS/SMB
2x10GbE

CES-1
NFS/SMB
2x10GbE

CES mount

EMS
2 x 10 GbE
2 x 56 Gb IB

GS4S NSD 1
AFM GW
2 x 40 GbE
4 x 56Gb IB

GS4S NSD 2
AFM GW
2 x 40 GbE
4 x 56Gb IB

1 x 10 GbE
1 x 1 GbE

Standard rate detectors
or control system clients
NFS or SMB mount from CES

1 x 40 GbE

Remote cluster mount

High rate detectors
Per beamline cluster
Native GPFS (or ZeroMQ)

1 x 10 GbE /
blade

1 x 56 Gb IB /
blade

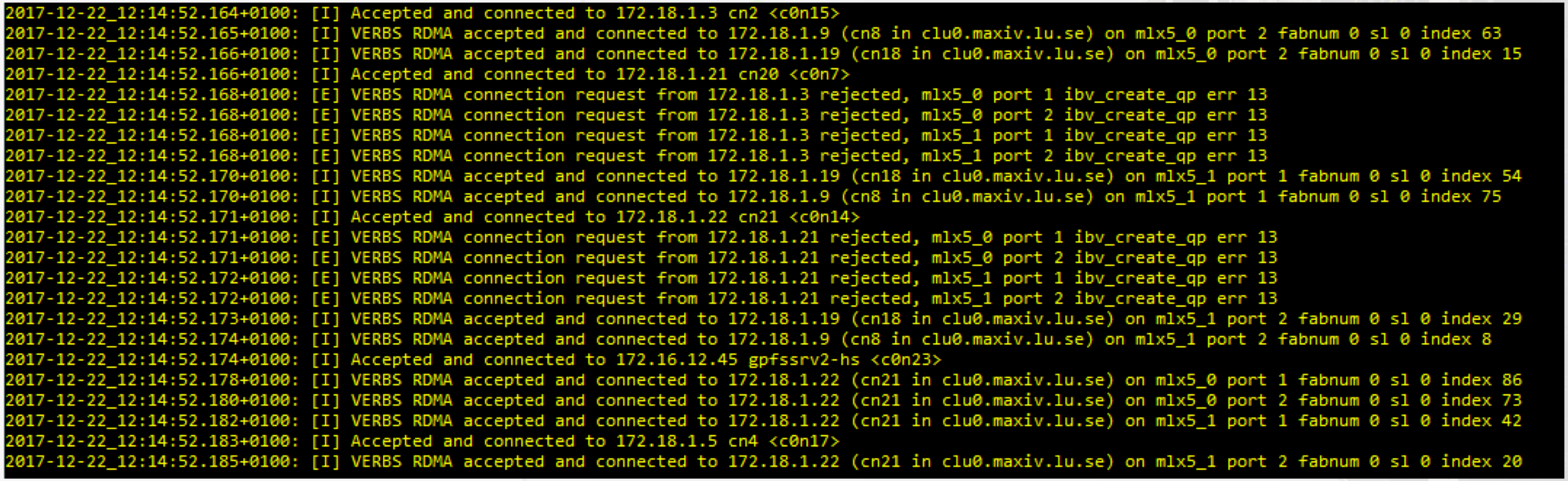
Remote cluster mount

16 node HPC Cluster
Separate Scale cluster

Spectrum Scale infrastructure at MAX IV

with some problems

Extract of mmfs.log from NSD node 2 of GS4S

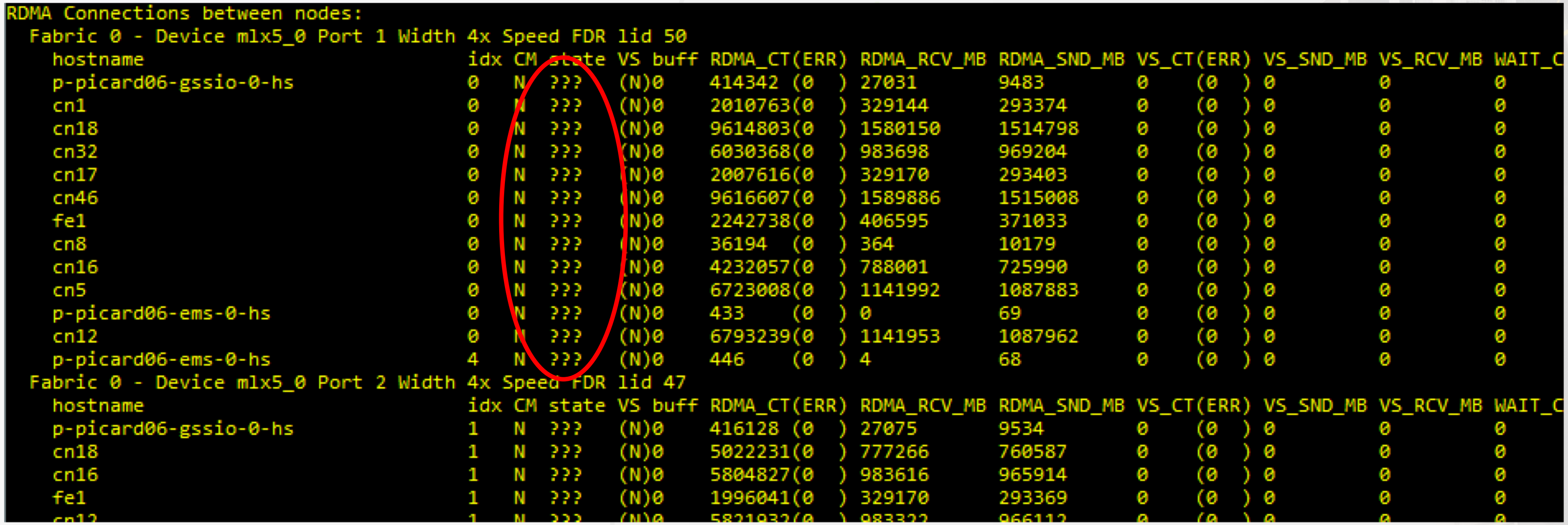


```
2017-12-22_12:14:52.164+0100: [I] Accepted and connected to 172.18.1.3 cn2 <c0n15>
2017-12-22_12:14:52.165+0100: [I] VERBS RDMA accepted and connected to 172.18.1.9 (cn8 in clu0.maxiv.lu.se) on mlx5_0 port 2 fabnum 0 sl 0 index 63
2017-12-22_12:14:52.166+0100: [I] VERBS RDMA accepted and connected to 172.18.1.19 (cn18 in clu0.maxiv.lu.se) on mlx5_0 port 2 fabnum 0 sl 0 index 15
2017-12-22_12:14:52.166+0100: [I] Accepted and connected to 172.18.1.21 cn20 <c0n7>
2017-12-22_12:14:52.168+0100: [E] VERBS RDMA connection request from 172.18.1.3 rejected, mlx5_0 port 1 ibv_create_qp err 13
2017-12-22_12:14:52.168+0100: [E] VERBS RDMA connection request from 172.18.1.3 rejected, mlx5_0 port 2 ibv_create_qp err 13
2017-12-22_12:14:52.168+0100: [E] VERBS RDMA connection request from 172.18.1.3 rejected, mlx5_1 port 1 ibv_create_qp err 13
2017-12-22_12:14:52.168+0100: [E] VERBS RDMA connection request from 172.18.1.3 rejected, mlx5_1 port 2 ibv_create_qp err 13
2017-12-22_12:14:52.170+0100: [I] VERBS RDMA accepted and connected to 172.18.1.19 (cn18 in clu0.maxiv.lu.se) on mlx5_1 port 1 fabnum 0 sl 0 index 54
2017-12-22_12:14:52.170+0100: [I] VERBS RDMA accepted and connected to 172.18.1.9 (cn8 in clu0.maxiv.lu.se) on mlx5_1 port 1 fabnum 0 sl 0 index 75
2017-12-22_12:14:52.171+0100: [I] Accepted and connected to 172.18.1.22 cn21 <c0n14>
2017-12-22_12:14:52.171+0100: [E] VERBS RDMA connection request from 172.18.1.21 rejected, mlx5_0 port 1 ibv_create_qp err 13
2017-12-22_12:14:52.171+0100: [E] VERBS RDMA connection request from 172.18.1.21 rejected, mlx5_0 port 2 ibv_create_qp err 13
2017-12-22_12:14:52.172+0100: [E] VERBS RDMA connection request from 172.18.1.21 rejected, mlx5_1 port 1 ibv_create_qp err 13
2017-12-22_12:14:52.172+0100: [E] VERBS RDMA connection request from 172.18.1.21 rejected, mlx5_1 port 2 ibv_create_qp err 13
2017-12-22_12:14:52.173+0100: [I] VERBS RDMA accepted and connected to 172.18.1.19 (cn18 in clu0.maxiv.lu.se) on mlx5_1 port 2 fabnum 0 sl 0 index 29
2017-12-22_12:14:52.174+0100: [I] VERBS RDMA accepted and connected to 172.18.1.9 (cn8 in clu0.maxiv.lu.se) on mlx5_1 port 2 fabnum 0 sl 0 index 8
2017-12-22_12:14:52.174+0100: [I] Accepted and connected to 172.16.12.45 gpfssrv2-hs <c0n23>
2017-12-22_12:14:52.178+0100: [I] VERBS RDMA accepted and connected to 172.18.1.22 (cn21 in clu0.maxiv.lu.se) on mlx5_0 port 1 fabnum 0 sl 0 index 86
2017-12-22_12:14:52.180+0100: [I] VERBS RDMA accepted and connected to 172.18.1.22 (cn21 in clu0.maxiv.lu.se) on mlx5_0 port 2 fabnum 0 sl 0 index 73
2017-12-22_12:14:52.182+0100: [I] VERBS RDMA accepted and connected to 172.18.1.22 (cn21 in clu0.maxiv.lu.se) on mlx5_1 port 1 fabnum 0 sl 0 index 42
2017-12-22_12:14:52.183+0100: [I] Accepted and connected to 172.18.1.5 cn4 <c0n17>
2017-12-22_12:14:52.185+0100: [I] VERBS RDMA accepted and connected to 172.18.1.22 (cn21 in clu0.maxiv.lu.se) on mlx5_1 port 2 fabnum 0 sl 0 index 20
```

Spectrum Scale infrastructure at MAX IV

with some problems

mmdiag --network strangeness



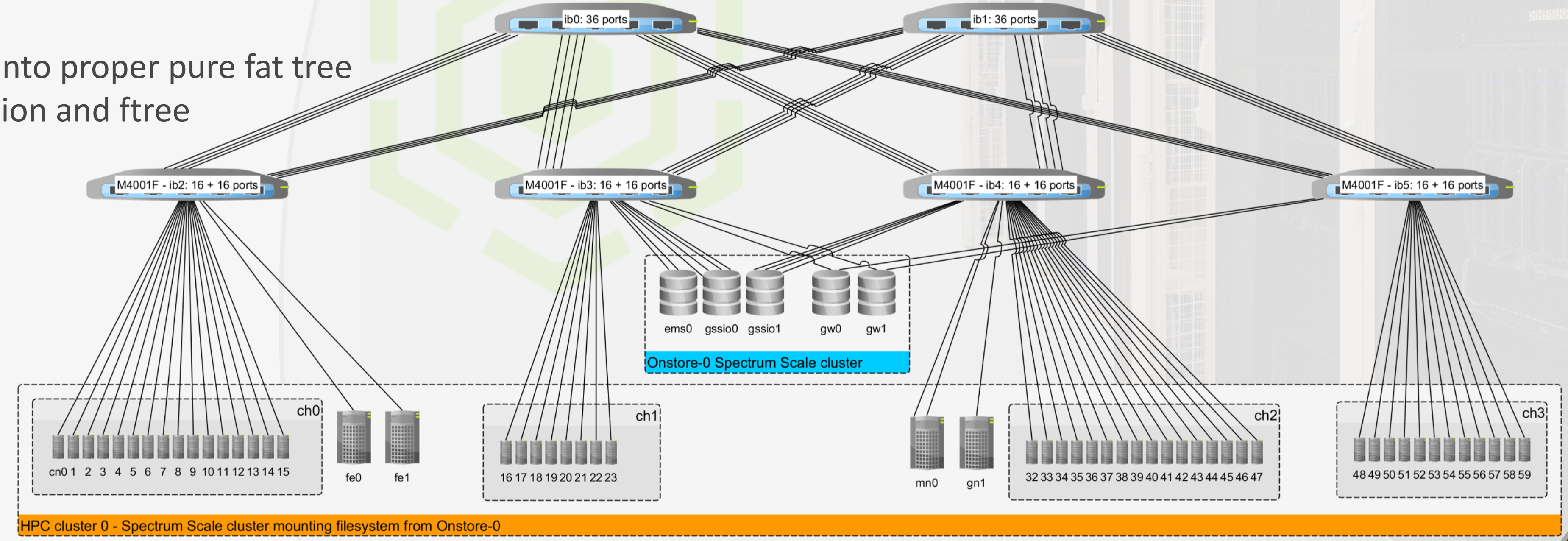
```
RDMA Connections between nodes:
  Fabric 0 - Device mlx5_0 Port 1 Width 4x Speed FDR lid 50
    hostname                            idx CM state VS buff RDMA_CT(ERR) RDMA_RCV_MB RDMA_SND_MB VS_CT(ERR) VS_SND_MB VS_RCV_MB WAIT_C
    p-picard06-gssio-0-hs               0   N  ???   (N)0    414342 (0  ) 27031       9483        0    (0  ) 0         0         0
    cn1                                 0   N  ???   (N)0    2010763(0  ) 329144      293374      0    (0  ) 0         0         0
    cn18                                0   N  ???   (N)0    9614803(0  ) 1580150     1514798     0    (0  ) 0         0         0
    cn32                                0   N  ???   (N)0    6030368(0  ) 983698      969204      0    (0  ) 0         0         0
    cn17                                0   N  ???   (N)0    2007616(0  ) 329170      293403      0    (0  ) 0         0         0
    cn46                                0   N  ???   (N)0    9616607(0  ) 1589886     1515008     0    (0  ) 0         0         0
    fe1                                 0   N  ???   (N)0    2242738(0  ) 406595      371033      0    (0  ) 0         0         0
    cn8                                 0   N  ???   (N)0    36194  (0  ) 364         10179       0    (0  ) 0         0         0
    cn16                                0   N  ???   (N)0    4232057(0  ) 788001      725990      0    (0  ) 0         0         0
    cn5                                 0   N  ???   (N)0    6723008(0  ) 1141992     1087883     0    (0  ) 0         0         0
    p-picard06-ems-0-hs                 0   N  ???   (N)0    433    (0  ) 0           69          0    (0  ) 0         0         0
    cn12                                0   N  ???   (N)0    6793239(0  ) 1141953     1087962     0    (0  ) 0         0         0
    p-picard06-ems-0-hs                 4   N  ???   (N)0    446    (0  ) 4           68          0    (0  ) 0         0         0
  Fabric 0 - Device mlx5_0 Port 2 Width 4x Speed FDR lid 47
    hostname                            idx CM state VS buff RDMA_CT(ERR) RDMA_RCV_MB RDMA_SND_MB VS_CT(ERR) VS_SND_MB VS_RCV_MB WAIT_C
    p-picard06-gssio-0-hs               1   N  ???   (N)0    416128 (0  ) 27075       9534        0    (0  ) 0         0         0
    cn18                                1   N  ???   (N)0    5022231(0  ) 777266      760587      0    (0  ) 0         0         0
    cn16                                1   N  ???   (N)0    5804827(0  ) 983616      965914      0    (0  ) 0         0         0
    fe1                                 1   N  ???   (N)0    1996041(0  ) 329170      293369      0    (0  ) 0         0         0
    cn12                                1   N  ???   (N)0    5821932(0  ) 983322      966112      0    (0  ) 0         0         0
```

Spectrum Scale infrastructure at MAX IV

Infiniband problems?

Checking IB links
Changing to Mellanox OFED subnet manager
Syncing OFED version of HPC with ESS
Updating all Ib cards and switches to latest firmware
Checking Ib error counters, etc.

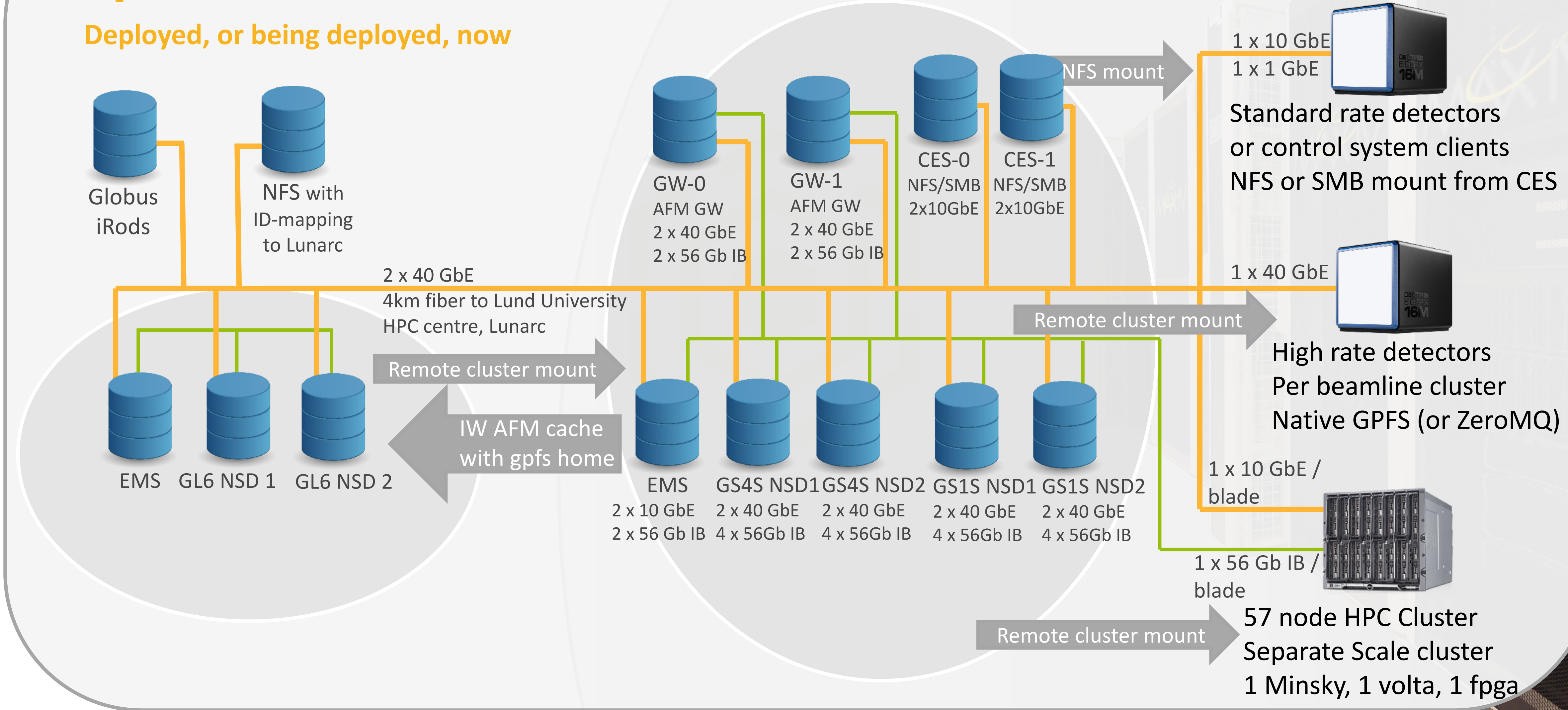
Rebuilding Infiniband into proper pure fat tree with 3:1 oversubscription and ftree routing algorithm:



ib0: 36 ports
ib1: 36 ports
M4001F - ib2: 16 + 16 ports
M4001F - ib3: 16 + 16 ports
M4001F - ib4: 16 + 16 ports
M4001F - ib5: 16 + 16 ports
ems0
gssio0
gssio1
gw0
gw1
Onstore-0 Spectrum Scale cluster
ch0
cn0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
fe0
fe1
ch1
16 17 18 19 20 21 22 23
mn0
gn1
ch2
32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
ch3
48 49 50 51 52 53 54 55 56 57 58 59
HPC cluster 0 - Spectrum Scale cluster mounting filesystem from Onstore-0

Spectrum Scale infrastructure at MAX IV

Deployed, or being deployed, now

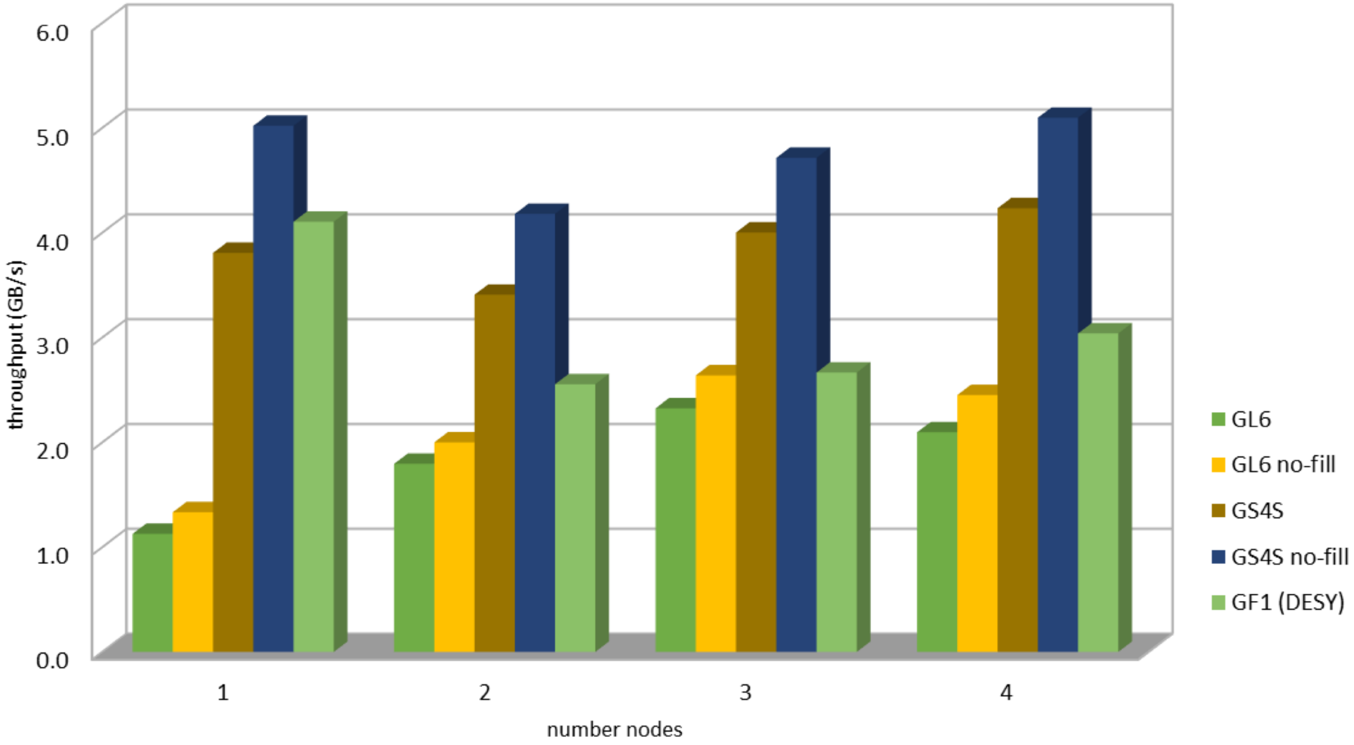


Standard rate detectors
or control system clients
NFS or SMB mount from CES

High rate detectors
Per beamline cluster
Native GPFS (or ZeroMQ)

57 node HPC Cluster
Separate Scale cluster
1 Minsky, 1 volta, 1 fpga

ESS GS4S vs GL6 vs GF1, workload simulation



The End

Thank you