Cluster of Research Infrastructures
for Synergies in Physics

SKA in WP19

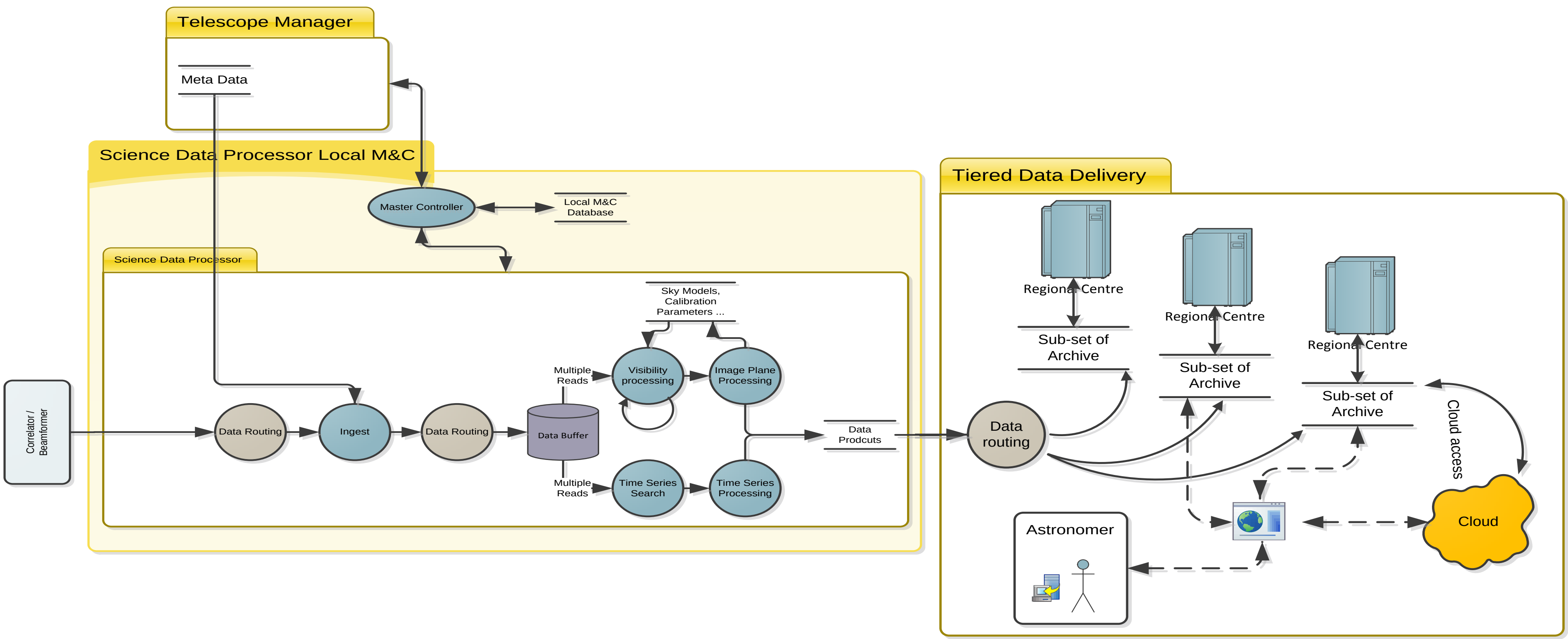
Iain Emsley

Rahim ‘Raz’ Lakhoo

David Wallom

3rd Annual Meeting

Science Data Processor



Consortium from: Australia, Canada, Chile, Germany, The Netherlands, New Zealand, Portugal, Spain, UK and USA

SDP Tasks and CRISP

- Compute platform
  - Input Handling
- Data Layer
  - Buffer
  - External interfaces
- Data delivery and tiered data model
  - Tiered data delivery model

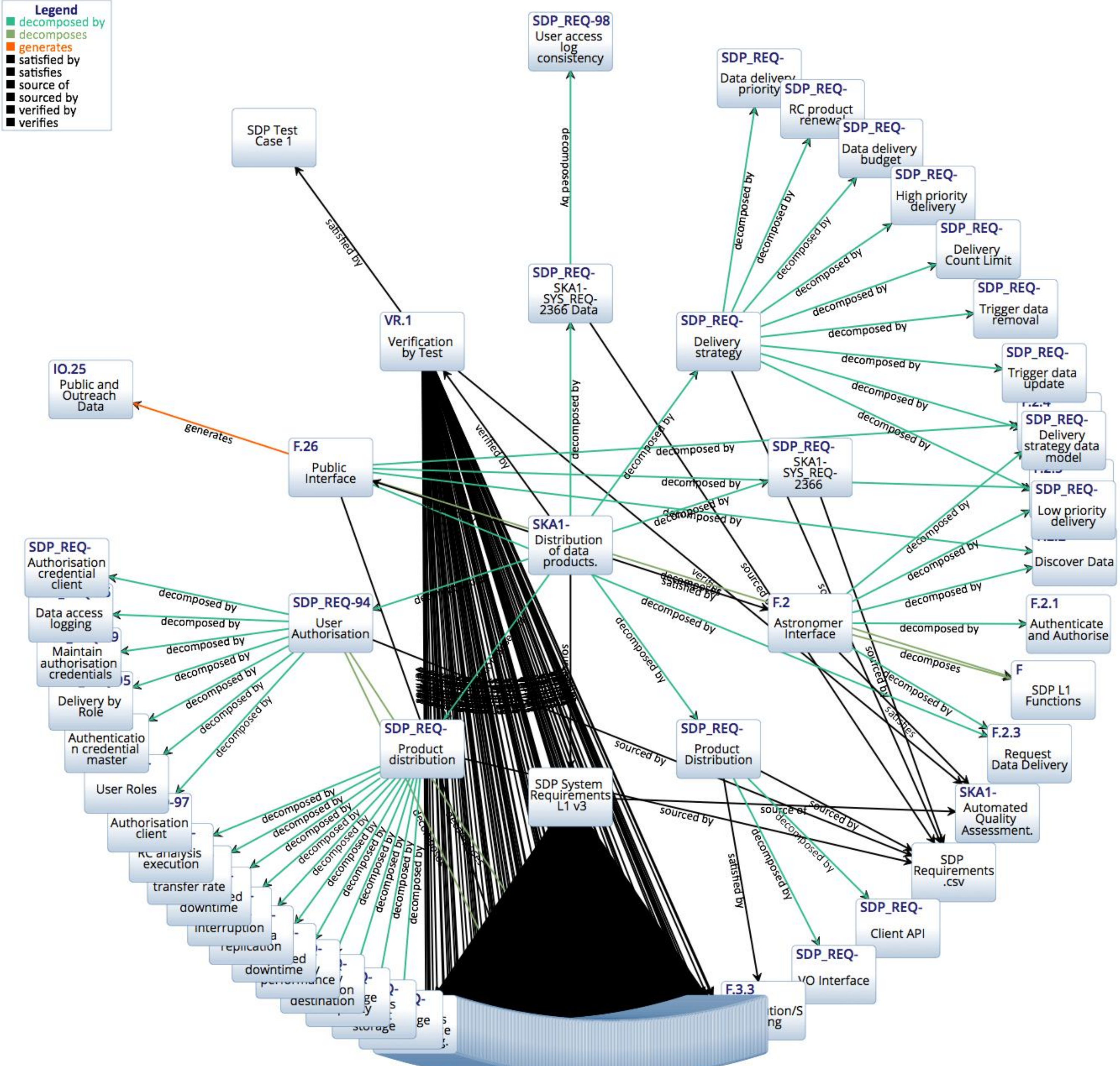
SKA WP 19 tasks

- Data Requirements
- Network Requirements
- Technology Survey
- Testbed

Data Requirements

- ~3.8 EB of processed data products over the project lifetime of SKA 1,
- 11 different types of data product types,
- Expected linear growth of data volume and data requests over the project lifetime,
- Data distributed to a number of regional centres*
- Access methods should include IVOA standards
- Support both SKA scientists and also provide public interfaces

- 19 published Science usecases
  - Range of data volumes and operational methodologies that will affect data volumes and transmission requirements


Legend
decomposed by
decomposes
generates
satisfied by
satisfies
source of
sourced by
verified by
verifies
SDP_REQ-98
User access log consistency
SDP Test Case 1
SDP_REQ-
SKA1-SYS_REQ-2366 Data
VR.1
Verification by Test
IO.25
Public and Outreach Data
F.26
Public Interface
SDP_REQ-
Delivery strategy
SDP_REQ-
Data delivery priority
SDP_REQ-
RC product renewal
SDP_REQ-
Data delivery budget
SDP_REQ-
High priority delivery
SDP_REQ-
Delivery Count Limit
SDP_REQ-
Trigger data removal
SDP_REQ-
Trigger data update
SDP_REQ-
Delivery strategy data model
SDP_REQ-
Low priority delivery
Discover Data
SDP_REQ-
SKA1-SYS_REQ-2366
SKA1-
Distribution of data products.
F.2
Astronomer Interface
F.2.1
Authenticate and Authorise
F
SDP L1 Functions
F.2.3
Request Data Delivery
SKA1-
Automated Quality Assessment.
SDP Requirements .csv
SDP_REQ-
Client API
SDP_REQ-
VO Interface
F.3.3
SDP_REQ-
Product Distribution
SDP System Requirements L1 v3
SDP_REQ-
Product distribution
SDP_REQ-94
User Authorisation
SDP_REQ-
Authorisation credential client
Data access logging
Maintain authorisation credentials
Delivery by Role
Authentication credential master
User Roles
Authorisation client
execution
transfer rate
downtime
interruption
replication
downtime
destination
storage
generates
satisfied by
verified by
decomposed by
decomposes
sourced by
source of
satisfies
satisfied by

Network Requirements

- ‘1Tbps link from each site would support SKA-1’

  *Not actually available currently from either site

- 1.4Tbps technology over existing infrastructure has been shown over ‘long’ distance
- Liaising with Signal and Data Transport consortium to connect to NREN and fibre providers (F2F in Manchester right now)

Delivery Tools Selection

- Level 3 – User Facing tools
  - e.g. CyberSKA
- Level 2 – Data distribution and management
  - e.g. iRODS, PhEDEx, NGAS
- Level 1 – Data transfer
  - e.g. GridFTP, RFT, FTS
- Level 0 – Lower level tools and protocols
  - Networking, messaging, RDMA

Example Review Criteria

- Scalability
  - Data demand
  - Data size
  - High-speed data transfers
  - Storage expandability
  - Architecture
- Data integrity
- Interoperability
- Data Discovery
- Management
- Fault tolerance
- Location aware
- Support and maintenance
- Software dependencies
- Open and standard protocols
- Software license
- Open standard security

Test Bed

2U chassis with x16 2.5 inch drive bays

Dual socket Intel E5-2690 CPUs, @2.9GHz (3.8GHz turbo), 135W TDP, 20MB Cache, 8GT/s QPI, Quad memory channel (Max. 51.2GB/s)

X9DRW-3LN4F+ Supermicro motherboard, BIOS version 3.00

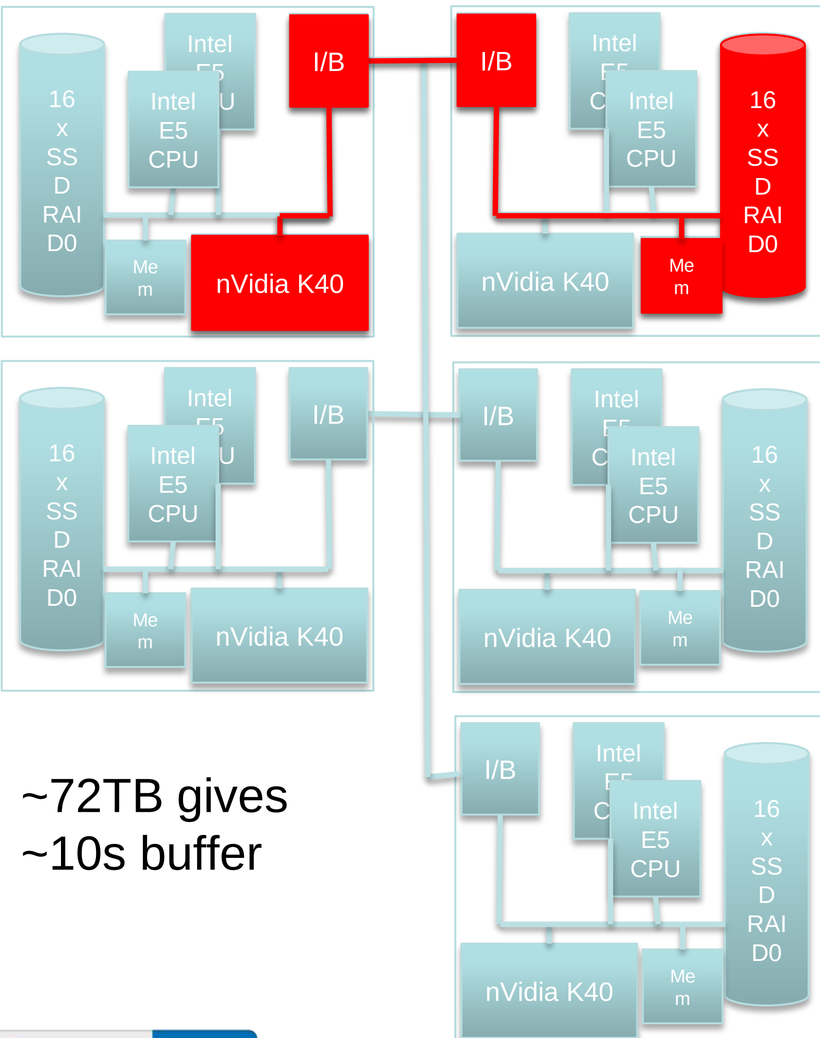
64GiB (8GiB x 8) ECC DDR3 1600MHz CL11 Single rank RAM @1.5V

Adaptec 71605 16 internal ports RAID card, PCIe 3.0 (x8), Mini-HD SAS, with 1GB DDR3 Cache
Sixteen 2.5 inch 128GB SATA3/6Gb OCZ Vector MLC NAND SSD's, ~2TB per node (RAID 0).
Mellanox FDR Infiniband 112Gbit/s (dual ports@56Gbit/s) Connect-IB PCIe 3.0 (x16)
Quad port Intel i350 GbE
GPUs include NVIDIA K10 (PCIe 3.0) and K40 (PCIe 3.0)
Dual 920W redundant PSUs

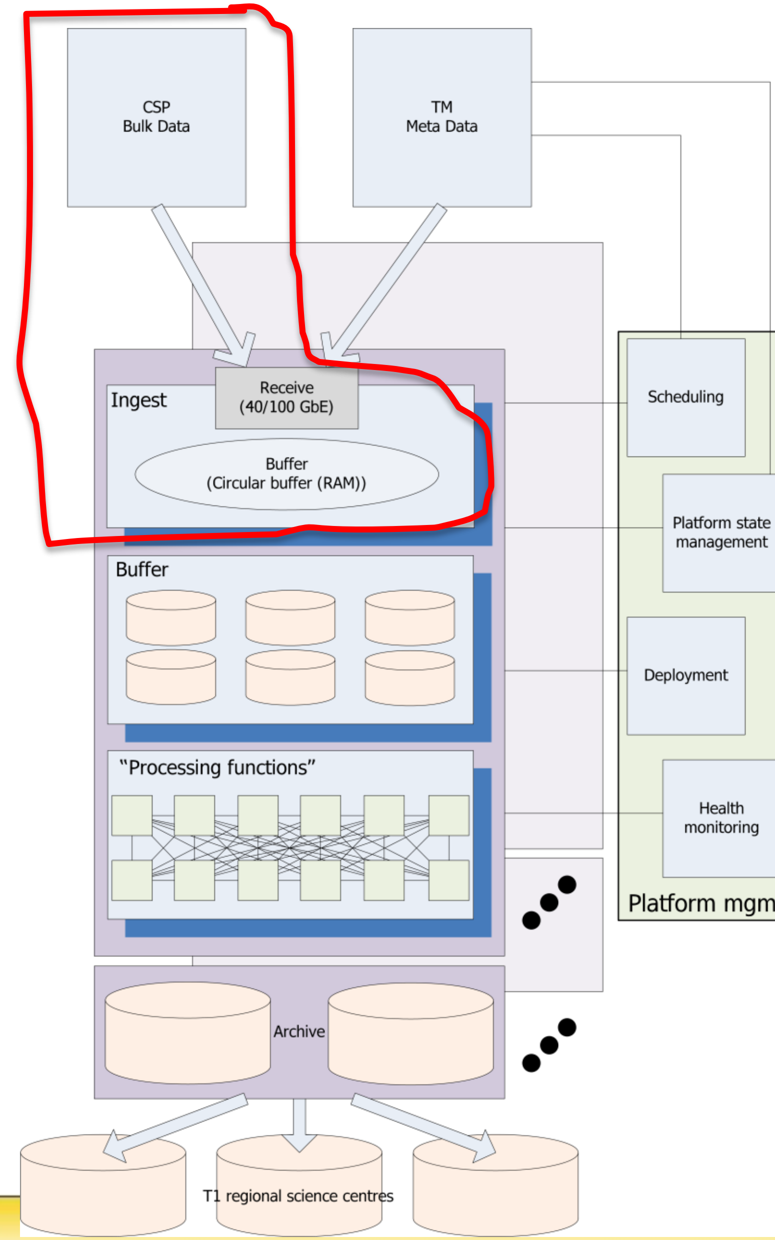
Test Harness

- Aims are re-usability and repeatability
- Installs PXE boot images of Debian, Centos with common settings
  - Ubuntu is in planning stage
- Puppet configuration management
  - Installs common packages, and
  - Software such as OSKAR and casacore
- Shared storage space for results

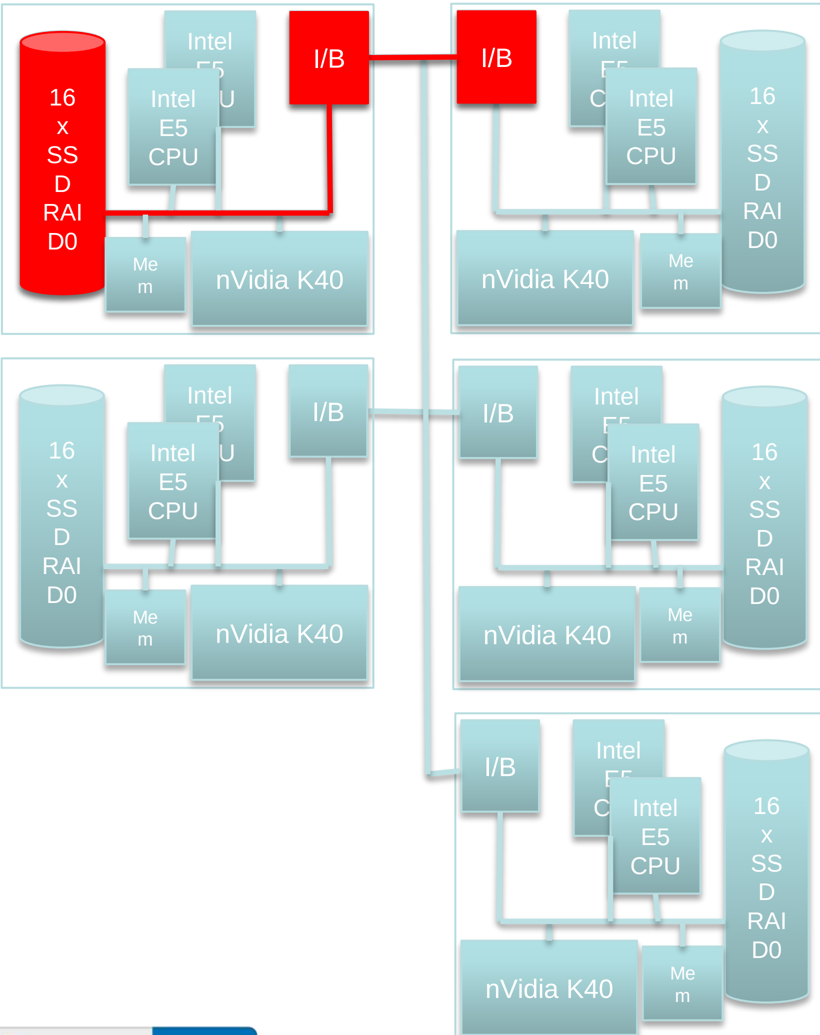
COMP.INP/DATA.BUF

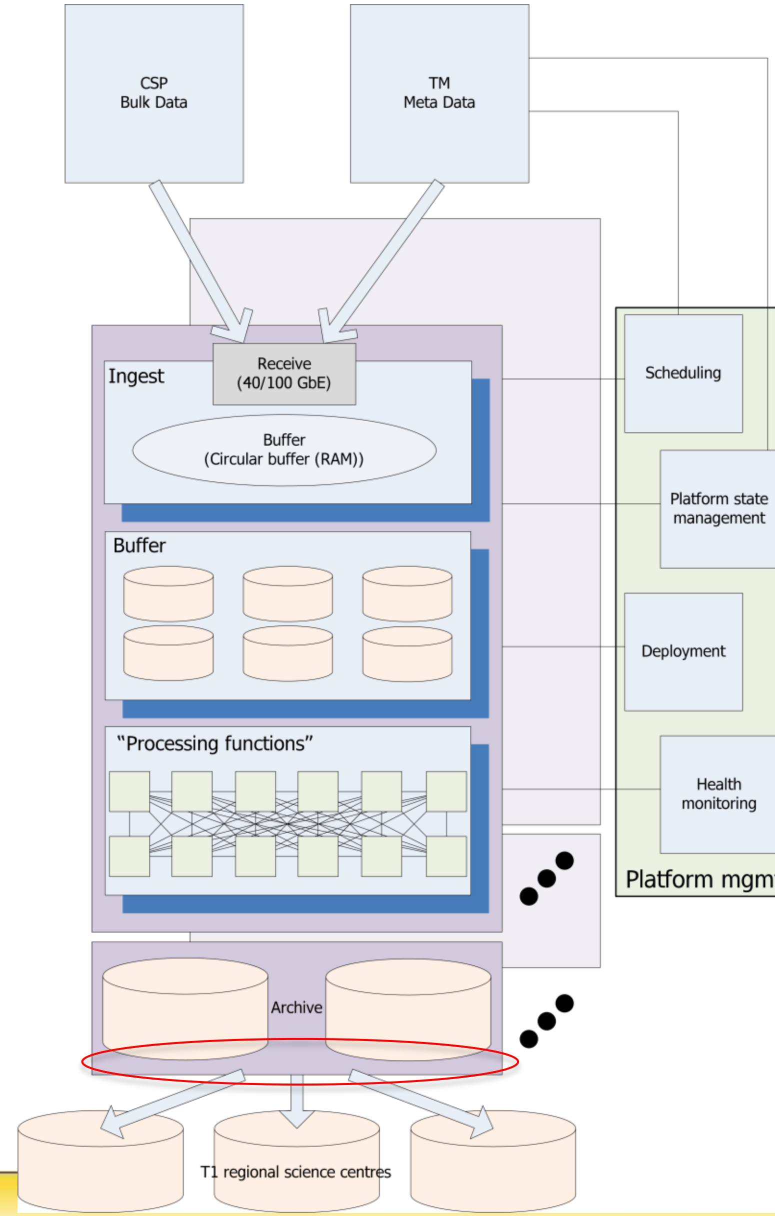


~72TB gives
~10s buffer

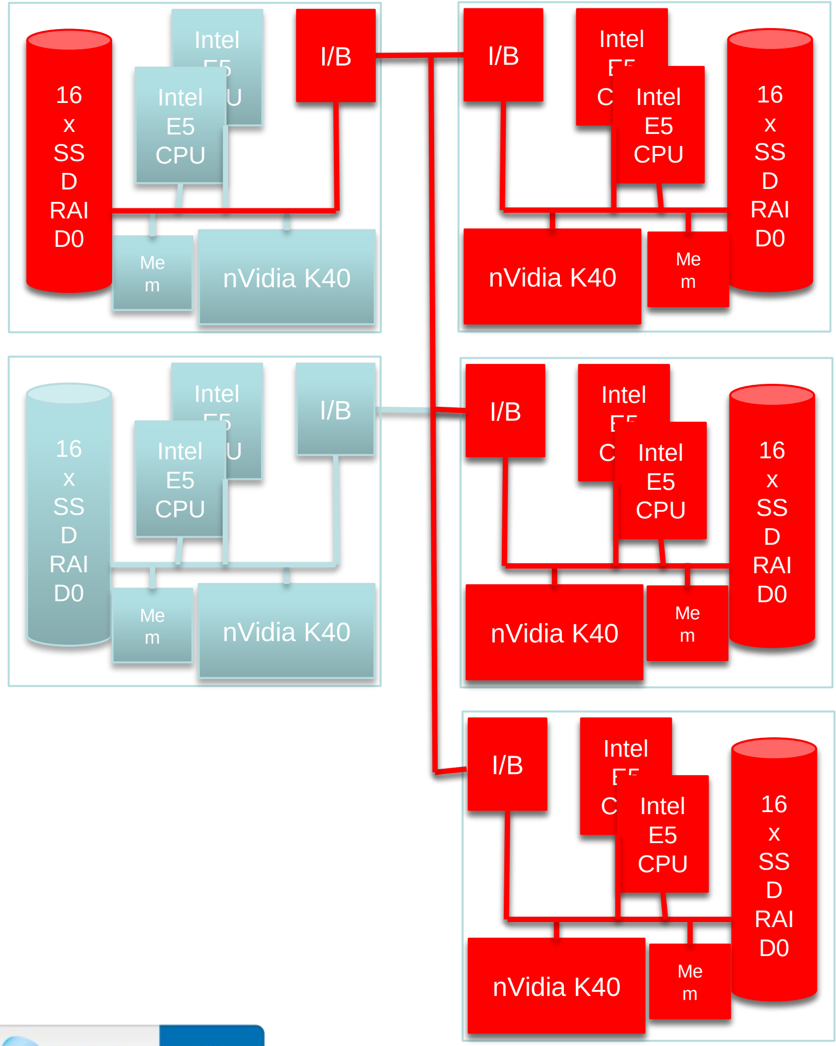


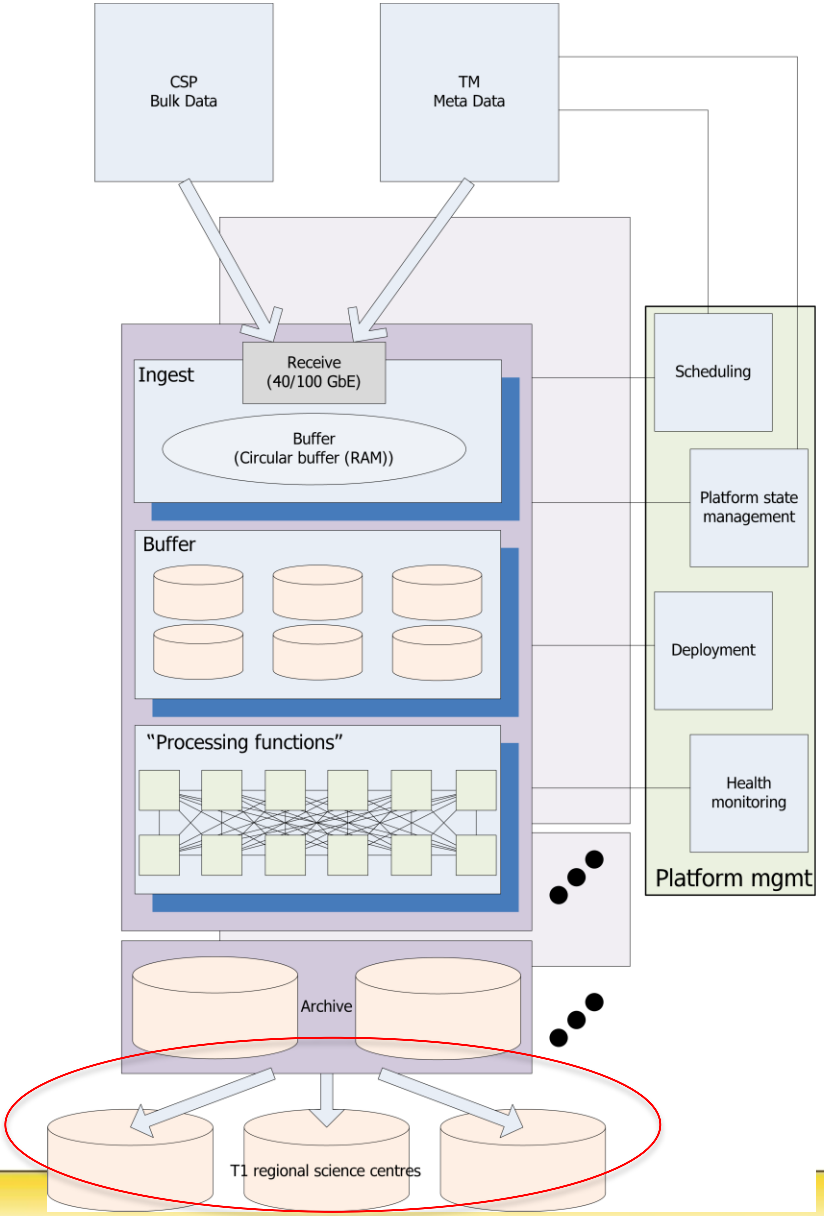
DATA.EXT


16 x SSD RAID0
Intel E5 CPU
Intel E5 CPU
I/B
Mem
nVidia K40
I/B
Intel E5 CPU
16 x SSD RAID0
nVidia K40
Mem
16 x SSD RAID0
Intel E5 CPU
I/B
Mem
nVidia K40
I/B
Intel E5 CPU
16 x SSD RAID0
nVidia K40
Mem
I/B
Intel E5 CPU
16 x SSD RAID0
nVidia K40
Mem

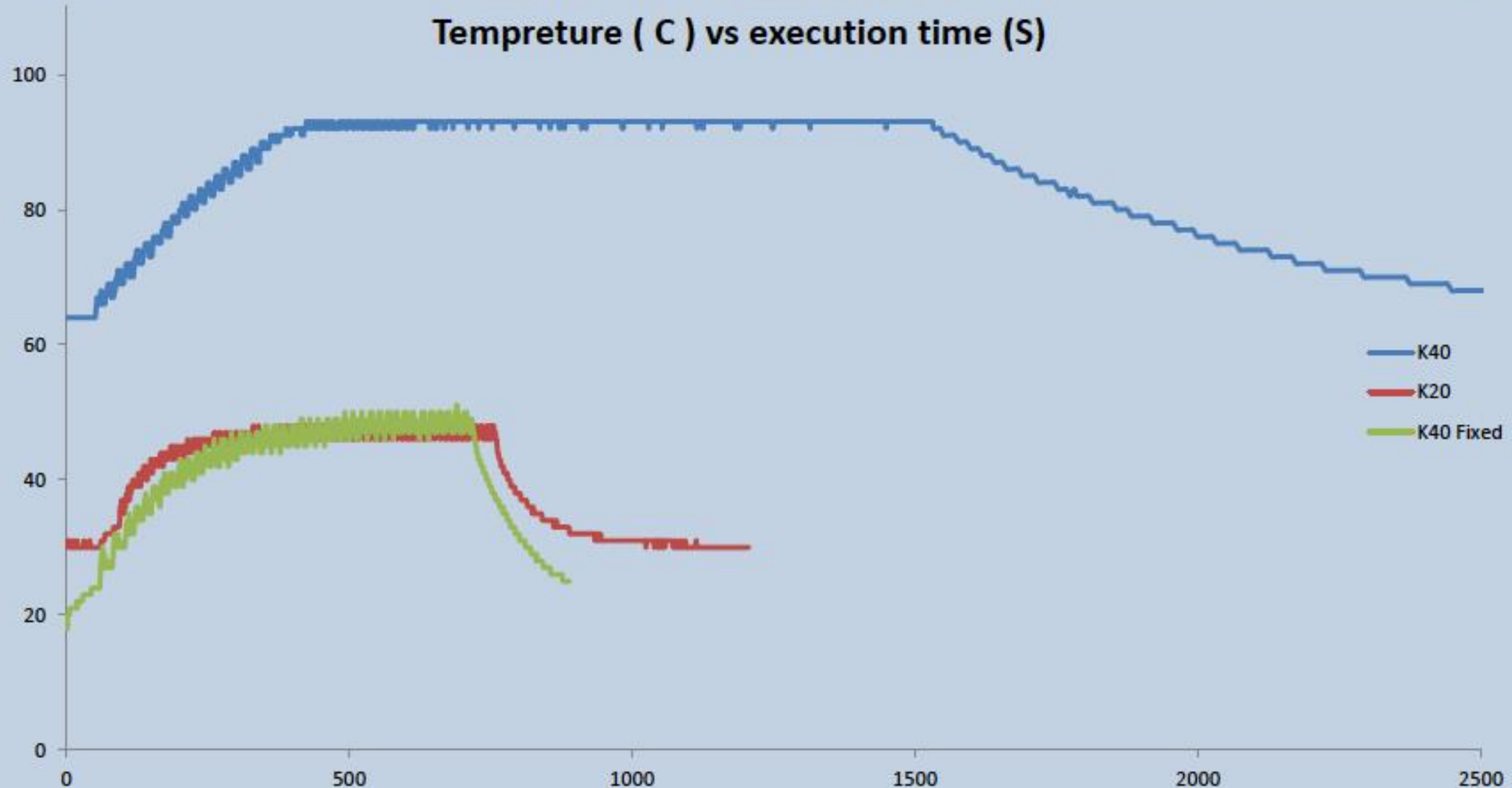

CSP
Bulk Data
TM
Meta Data
Ingest
Receive
(40/100 GbE)
Buffer
(Circular buffer (RAM))
Buffer
"Processing functions"
Archive
T1 regional science centres
Scheduling
Platform state management
Deployment
Health monitoring
Platform mgmt

DATA DELIVERY





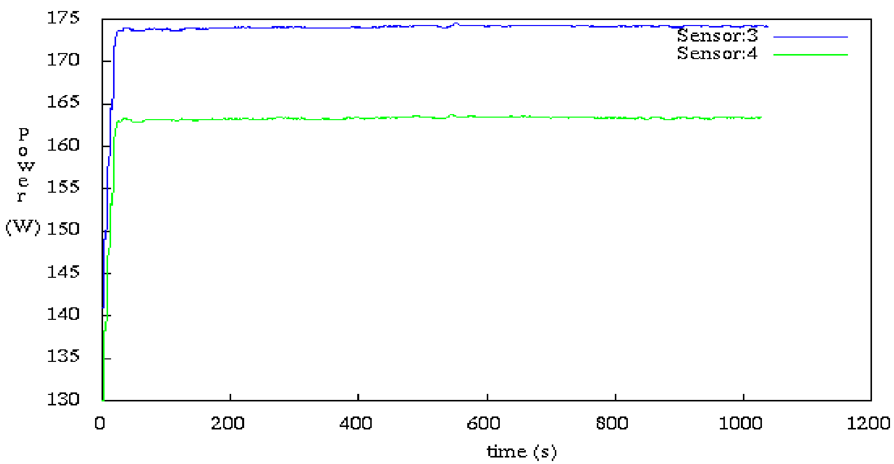
Hardware Tuning…

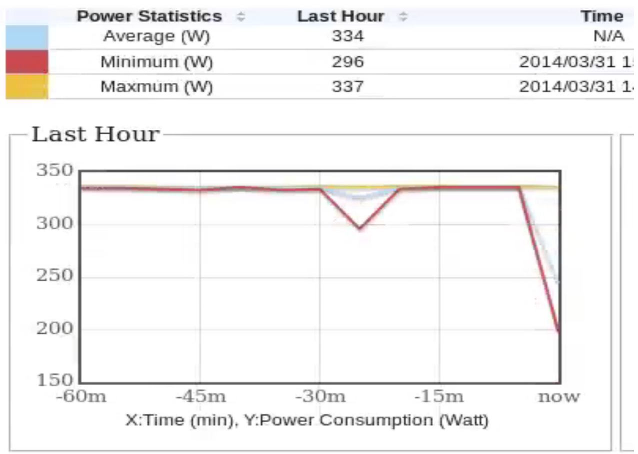


With Rahim (Raz) Lakhoo rahim.lakhoo@oerc.ox.ac.uk

Initial Energy Profiling

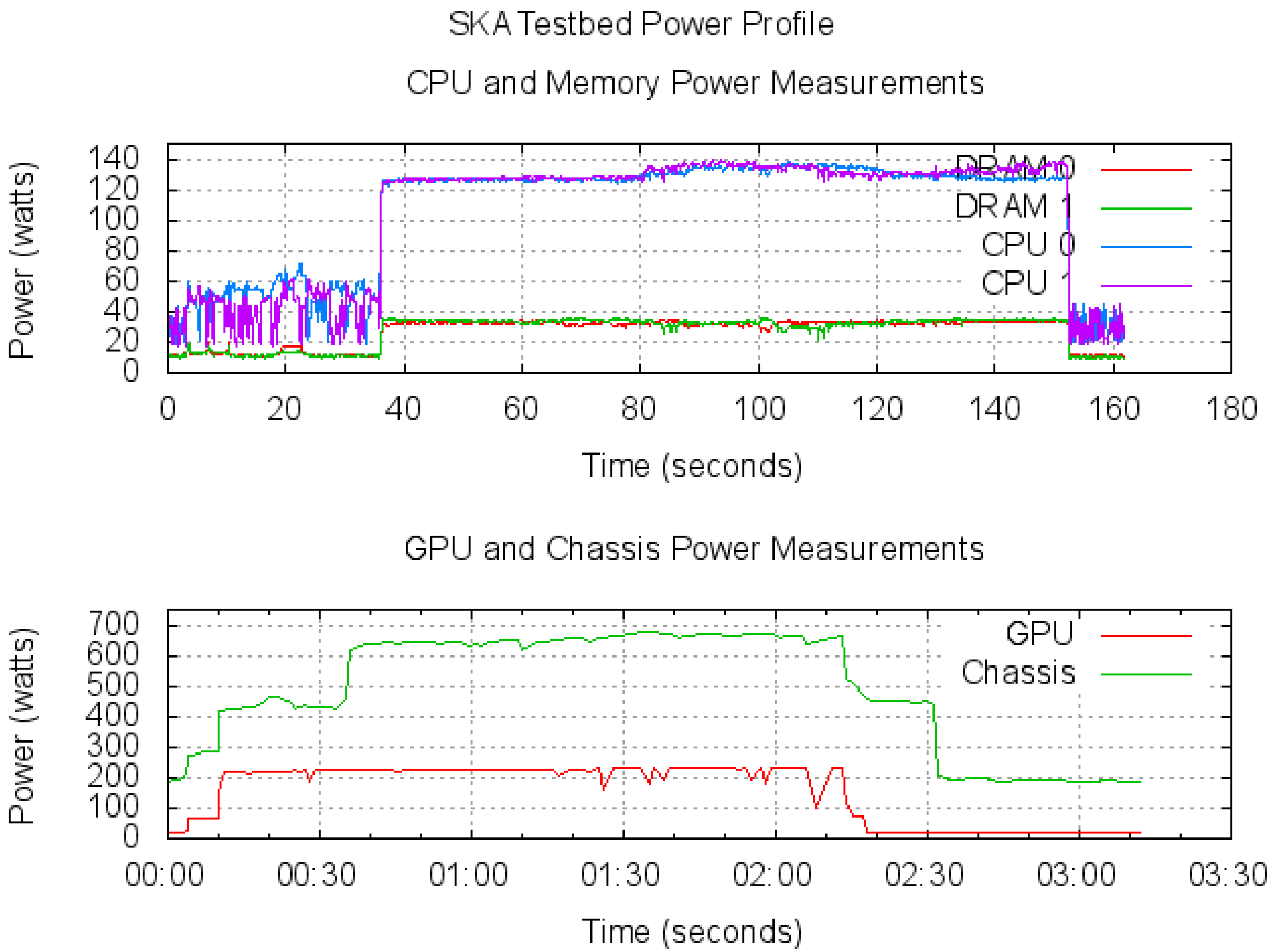
- Initial profiling using EMPPACK package with external Watt's Up Pro meters
- Testbed chassis power monitoring shows the same power usage, but with less resolution





Both power readings show approx. 338 watts - basic power device verification

Enhanced Power Monitoring

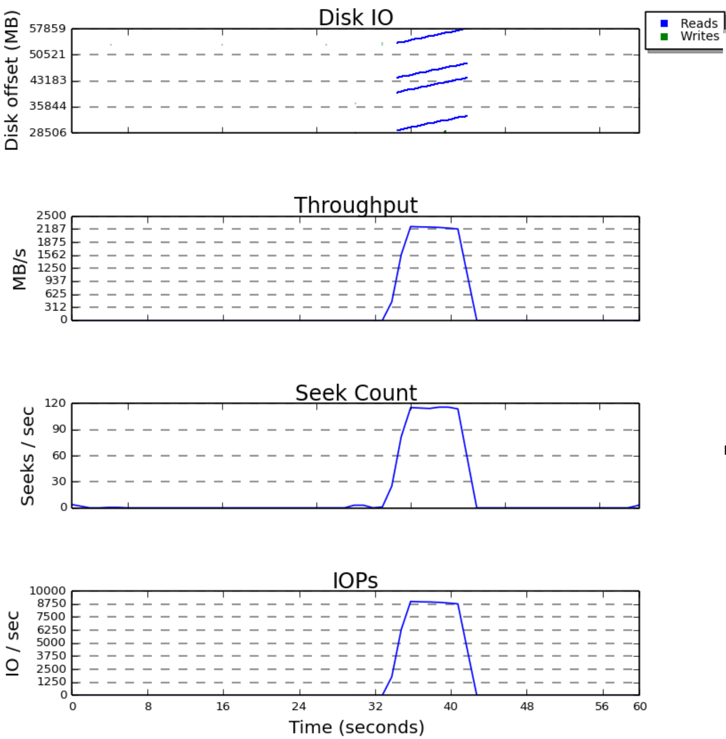


Use RAPL, IPMI and NVML, to give CPU and DRAM, Chassis and GPU power readings.

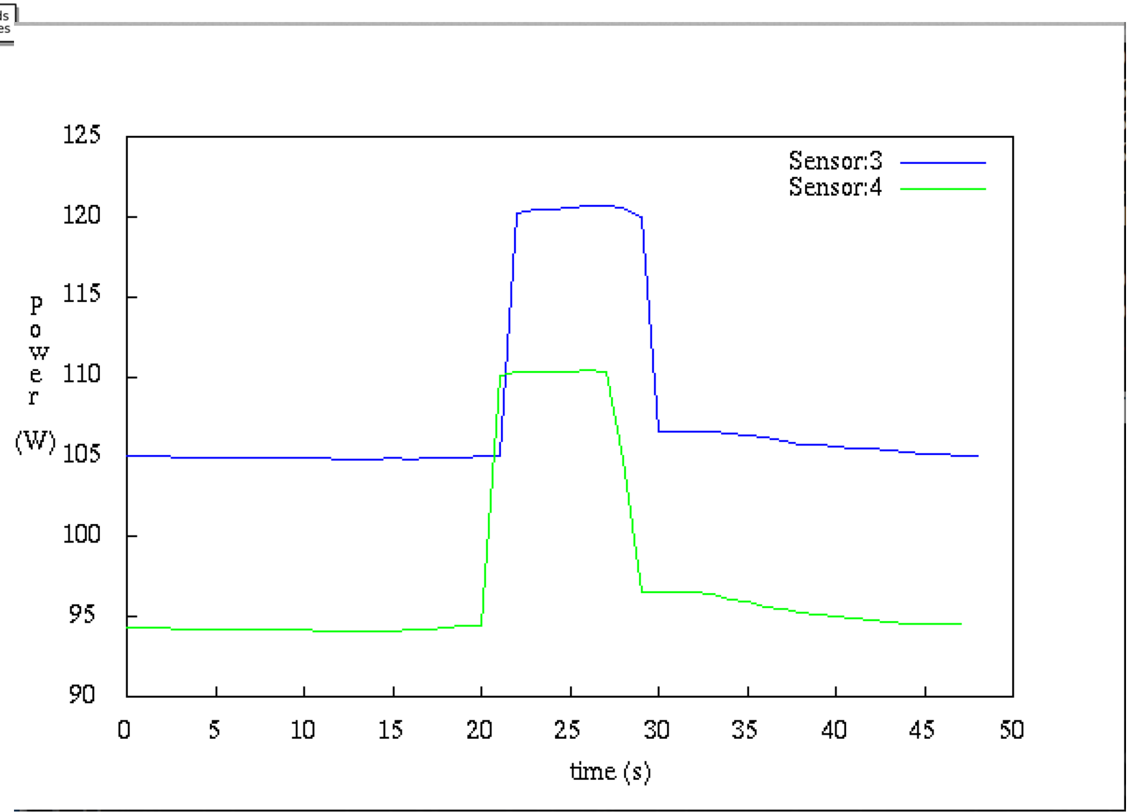
Developing Enhanced Application Profiling

- Low overhead profiling tools and methods,
- Capture power and performance for NICs, CPUs, GPUs (inc. GPUDirect), I/O devices, and Memory,
- Method needs to scale from small (i.e. ARM) to large systems (i.e. compute node).
- Kernel-level tool chains and tracers:
  - I/O profiling/tracing and replay has been done with a 2-5% overhead.
- Aiming for a ‘one stop shop’ for metric gathering

Profiling with Power



Avg Seeks/s	Avg MB/s	Avg IO/s	Run time (s)
14.49	274.16	1096.95	60.16

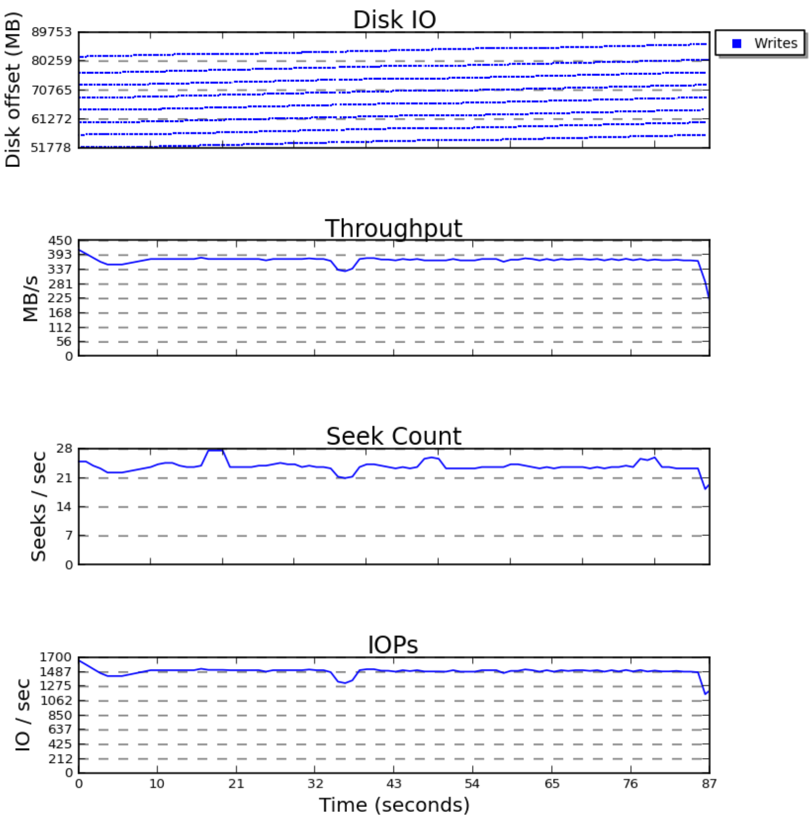


Average - 2100MB/s

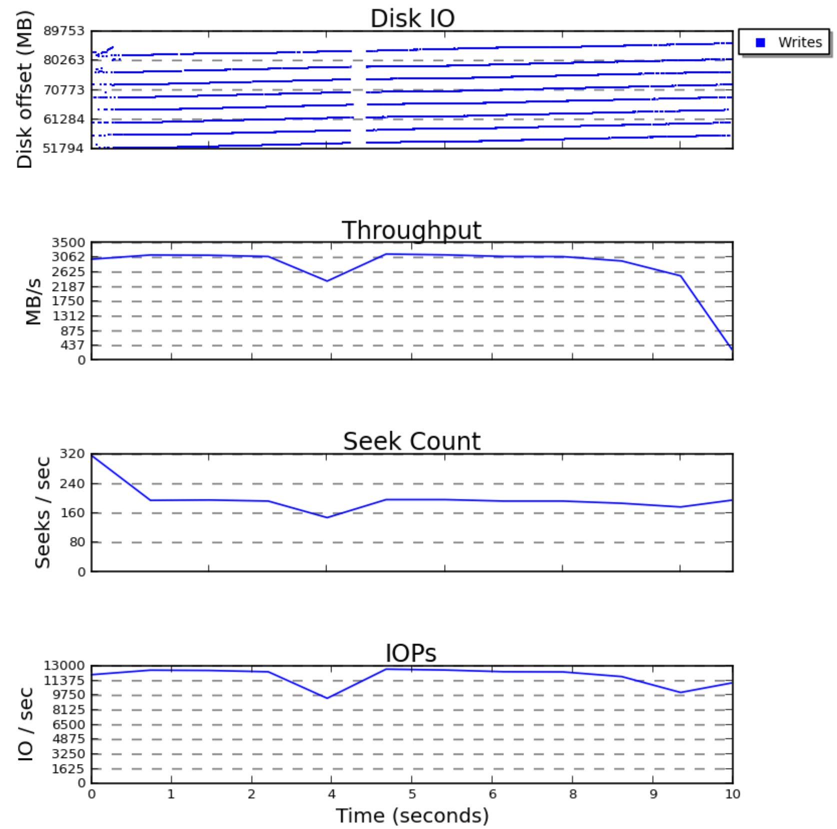
Approx. 230 watts power

Please ignore time scale on the right graph

Data Write

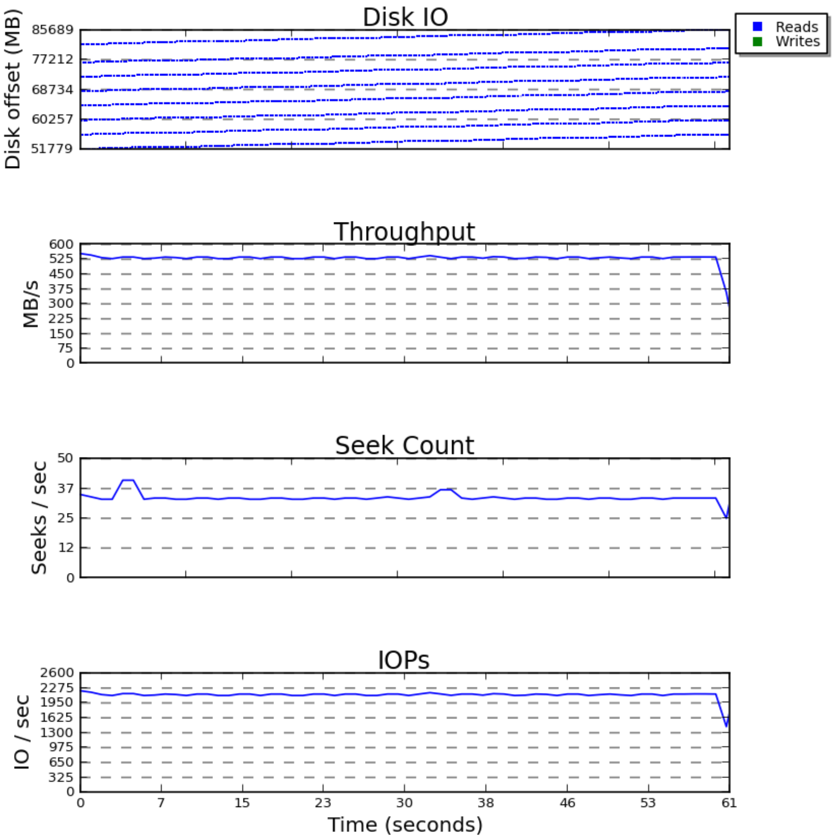


Avg Seeks/s	Avg MB/s	Avg IO/s	Run time (s)
23.79	373.4	1493.96	87.67

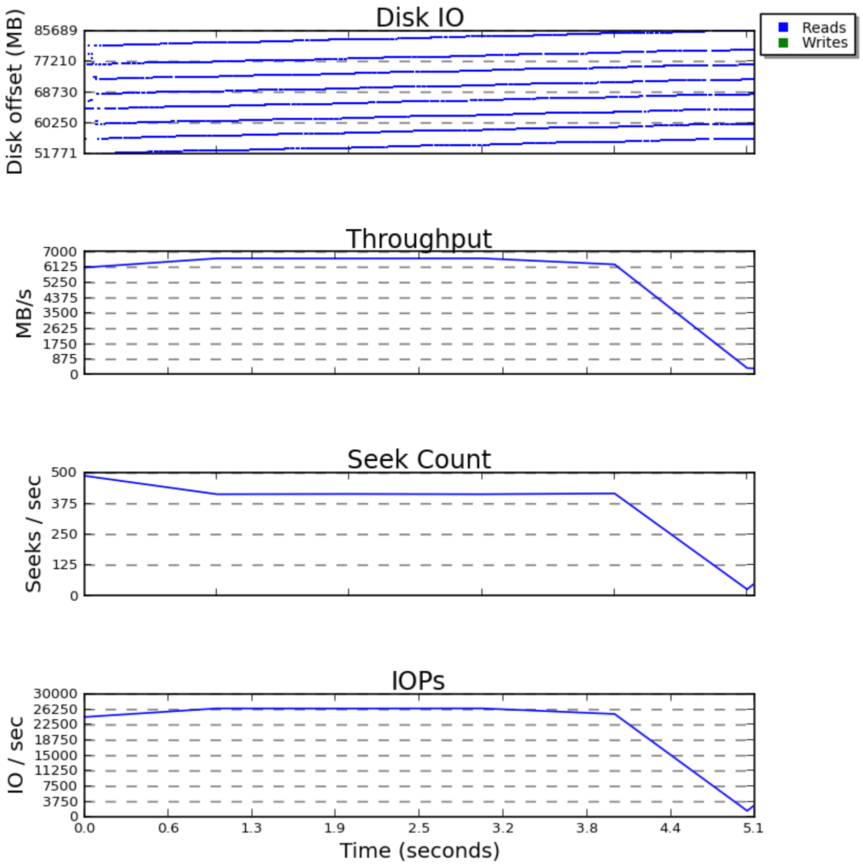


Avg Seeks/s	Avg MB/s	Avg IO/s	Run time (s)
201.33	3000.41	12003.87	10.89

Data Read



Avg Seeks/s	Avg MB/s	Avg IO/s	Run time (s)
33.83	533.51	2134.61	61.34



Avg Seeks/s	Avg MB/s	Avg IO/s	Run time (s)
430.19	6460.91	25847.86	5.06

Future Plans

- Point to point tests between SDP DELIV members on selected tools
- Configuration of protocols for lightpath setup to host countries
- Publish analysis of delivery tools
- Completing all workpackge L2/L3 requirements decomposition
- Profiling LOFAR and MWA algorithms
- Hardware FLOP counters to compliment hardware power counters
- Working with Mellanox for NIC power measurements
- GPUDirect configuration and integration in testbed
- Investigating BFS CPU scheduler and Kernel tweaks