iVEC Site Report

Andrew Elwell

Andrew.Elwell@ivec.org

ivec.org iVEC Advancing science through supercomputing

About iVEC

- 10+ years of energising research in and uptake of supercomputing, data and visualisation in WA and beyond
- Five Partners
  - CSIRO
  - Curtin University
  - Edith Cowan University
  - Murdoch University
  - The University of Western Australia
- Funded by
  - The Government of Western Australia
  - The Partners
  - The Commonwealth (Australian) Government

ivec.org iVEC Advancing science through supercomputing

iVEC Facilities and Expertise

40+ staff across five facilities, around Perth

- **CSIRO Pawsey Centre/ ARRC** – uptake, supercomputing , data, visualisation
- **Curtin University** – uptake and visualisation
- **Edith Cowan University** – uptake and visualisation
- **Murdoch University** – supercomputing
- **University of Western Australia** – supercomputing, uptake, visualisation
- Sites linked together by dedicated high-speed network, and to rest of world via AARNet

ivec.org iVEC Advancing science through supercomputing

SKA Location in Australia

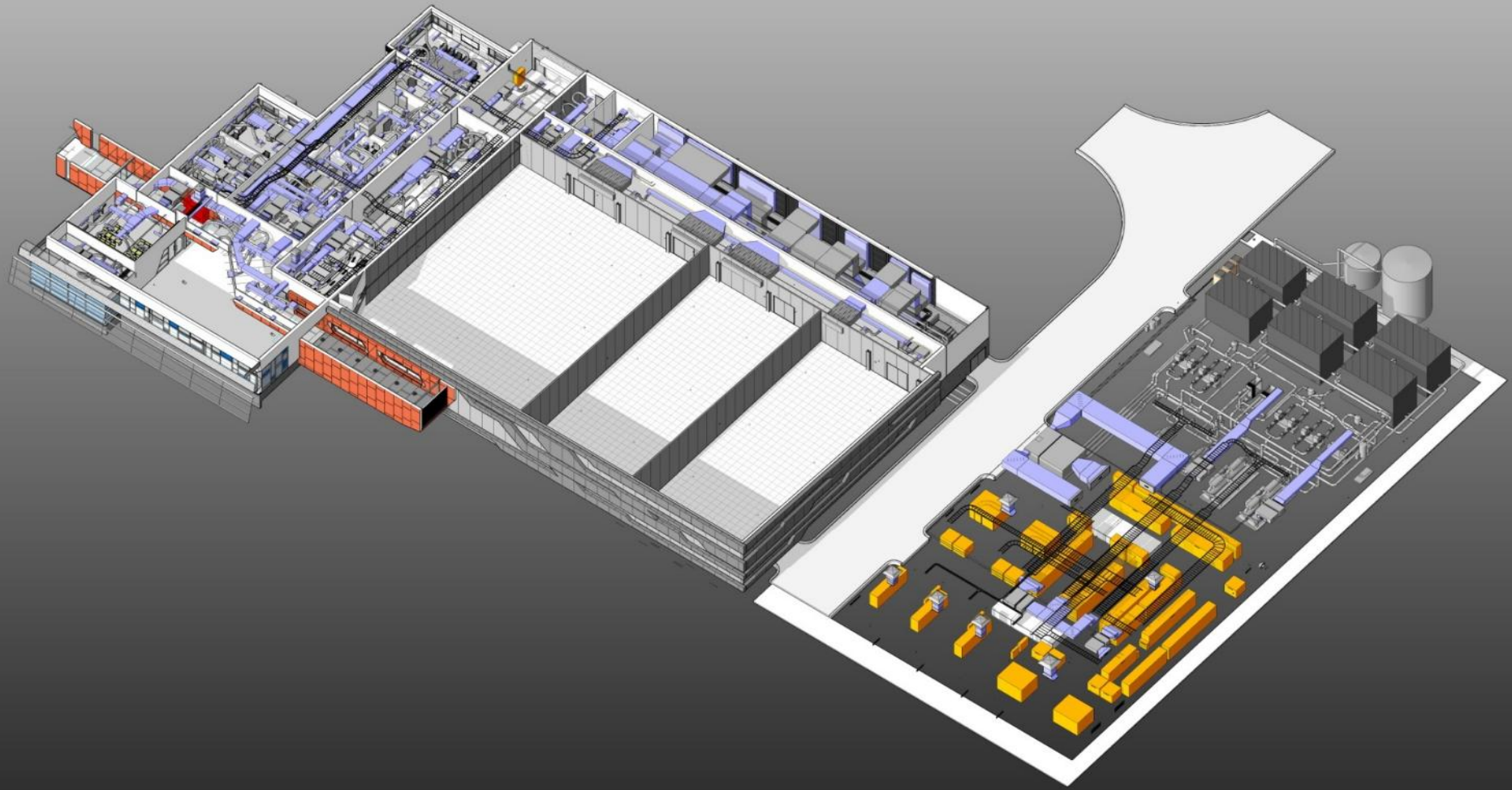
- Murchison Radio Observatory (MRO) (Shire of Murchison)
  - 41,172 km^2 radio-quiet desert region
  - Population never more than 120
  - For comparison …
    - Switzerland - area 41,285 km^2, pop: 8,014,000
    - Netherlands - area 41,543 km^2, pop: 16,789,000
  - Limited infrastructure for power, cooling, people
- Strike balance between *essential* data processing on-site and shipping data elsewhere

ivec.org iVEC Advancing science through supercomputing

Pawsey Centre

- In May 2009, Australian Government chose iVEC to establish and manage $80M AUD Pawsey Supercomputing Centre
- World-class petascale facility
  - to prepare for challenges of computing and data-processing for Square Kilometre Array (SKA)
- Provide significant advantage for Australian researchers
  - radio astronomy, geosciences, nanotechnology, physical sciences, marine science, engineering, bioinformatics, …

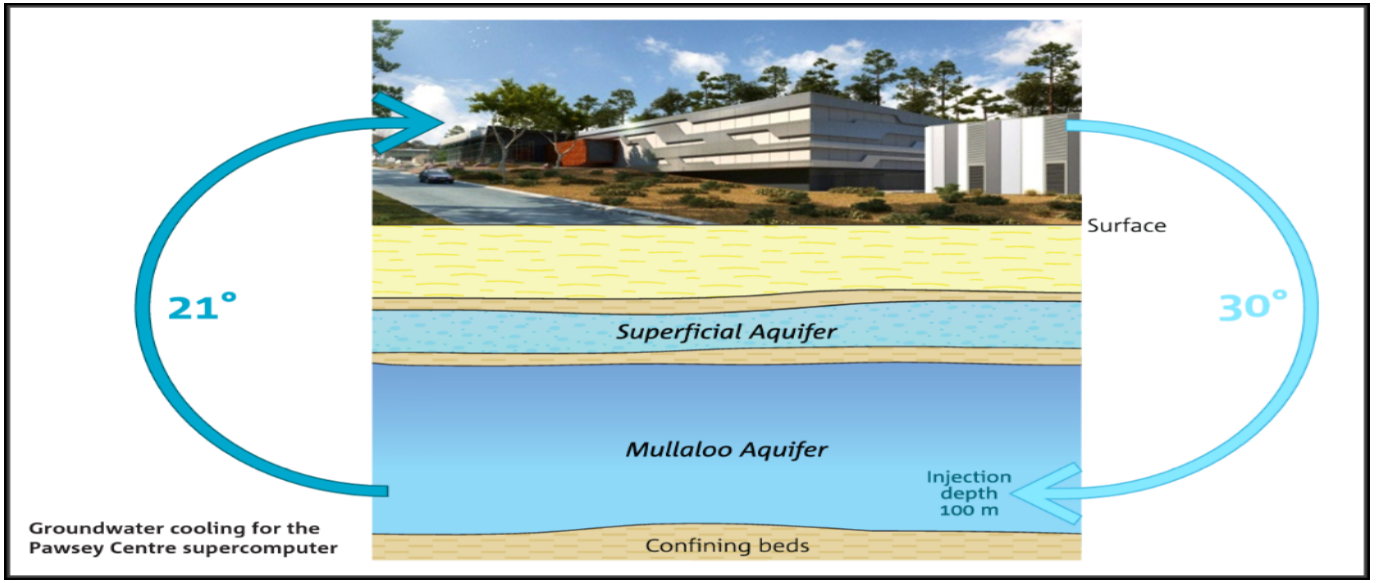
ivec.org iVEC Advancing science through supercomputing



Pawsey Centre Green Initiatives

ivec.org iVEC Advancing science through supercomputing

Ground-Water Cooling System



- CSIRO Sustainable Energy for the SKA project
- Closed-loop circuit provides cooling for supercomputers without need for chillers or cooling towers
- Estimated saving of 14M litres potable water per year

ivec.org iVEC Advancing science through supercomputing

Magnus

- 1488 Xeon E5-2690 v3 @ 2.60GHz Nodes
- 35000+ cores (1488*2*12)
- 3.5 PB Sonexion 1600 Lustre /scratch
- 850 TB site wide Lustre /group
- 14 TB site wide NFS /home
- 2 Data mover nodes
- 2 CDL (eslogin) nodes
- 12h max wallclock
  can have reservations for longer

ivec.org iVEC Advancing science through supercomputing

Galaxy

- 472 E5-2690 v2 @ 3.00GHz nodes
- 9440 Cores
- 64 K20x GPU nodes
- 1.3 PB Sonexion 1600 Lustre /scratch2
- 850 TB Lustre /group
- 14 TB NFS /home
- 16 Ingest (10GbE to MRO) nodes
- 2 General Datamover nodes
- 2 CDL (esLogin) nodes

ivec.org iVEC Advancing science through supercomputing

Chaos

- Single Chassis Air Cooled XC30
- 2 Haswell, 2 Ivybridge, 1 GPU blades
- IB connection ‘in progress’
- Single CDL Login node
- Testing with topology.conf


```
xtprocadmin | grep n0 | awk '{print "SwitchName="$3 " Nodes=nid000["$1"-"$1+3"]"}'
```

Zythos

- UV 2000
- 24h Max wallclock

The Register®

Data Centre | Software | Networks | Security | Business | Hardware | Science | Bootnotes | Video | Forums

Servers | HPC | Cloud | Storage | Data Networking | Virtualisation | BOFH

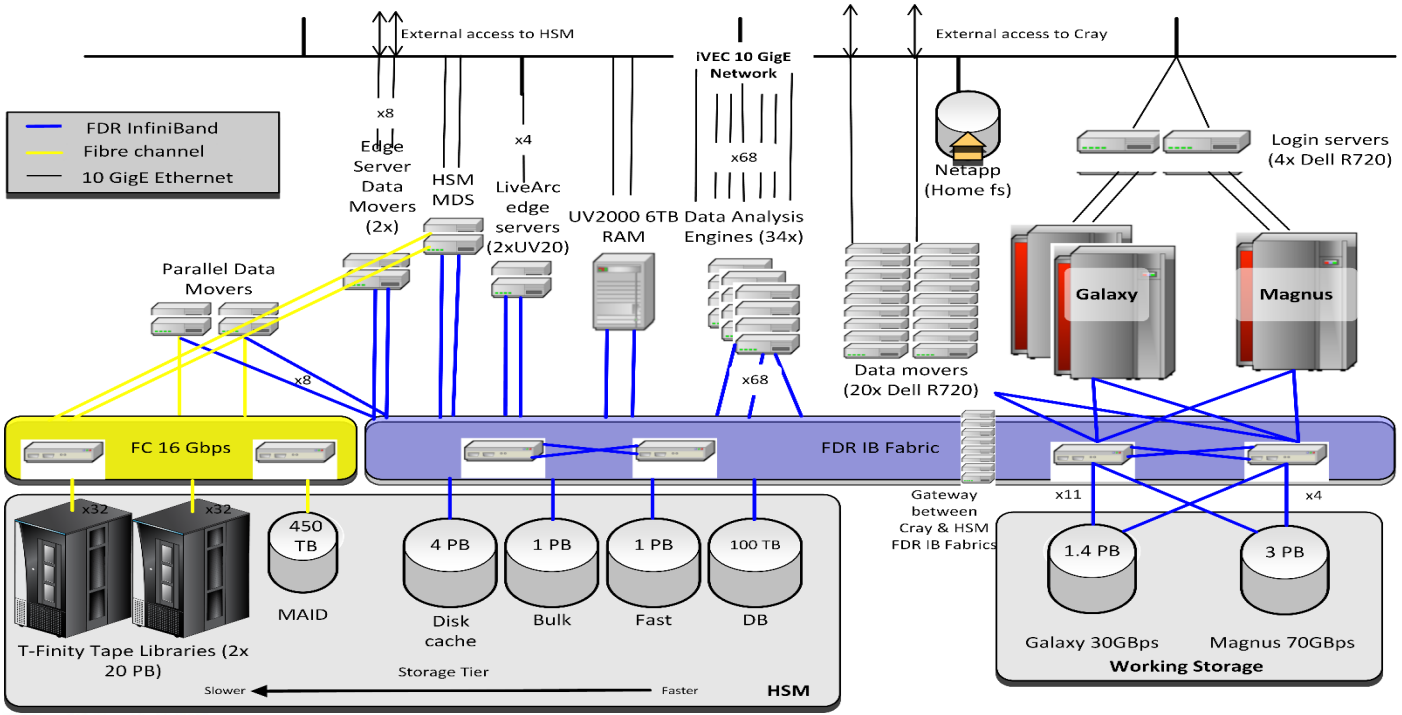
DATA CENTRE > HPC

What could you do with 264 Sandy Bridge cores and 6TB of RAM?

Perth supercomputing outfit iVEC offers researchers time on its Zythos BEAST

ivec.org iVEC *Advancing science through supercomputing*

System Arcitecture



ivec.org iVEC Advancing science through supercomputing

SLURM

- Chosen as Centre-wide batch scheduler
- New to all operations staff
- Slurmdbd on VM host, pointing to site MySQL server
- 2.6.6-2-ivec-1 and 2.6.9

Install on Crays

- ./configure finds /opt/cray on eslogin nodes
- Backported some bugfixes from later 2.6.x
- Using ALPS/Basil – Old API version
  LATEST gives xml offset error
- Run slurmctld on SDB nodes (routable IP addr)
- Slurmd's on both mom nodes
- Salloc wrapper script to ssh ${IVEC_CLUSTER}-int
- Usage reports in sreport is 50% low
- GRES GPU didn't seem to work (slurmd swallowed output)

Install on other systems

- Vanilla SLES11sp3 build on non-cray box
- cgroups for zythos
- Aliased qsub / qstat to informational message
- Set Features – Chipset, GPU type etc

- TODO – reorganise configs
  include ${IVEC_CLUSTER}.conf

Environment

- Module exports “pretty” variables

SINFO_FORMAT %9P %5a %8s %.10l %.6c %.6z %.7D %10T %N

SQUEUE_FORMAT %8i %8u %15a %.14j %.9B%.3t %.10r %.12S %.12e %.10L %.5D %.10Q

SQUEUE_ALL yes

SQUEUE_SORT -t,-e,-p

SLURM_TIME_FORMAT relative

JOBID	USER	ACCOUNT	NAME	EXEC_HOST	ST	REASON	START_TIME	END_TIME	TIME_LEFT	NODES	PRIORITY
119497	jwang	partner769	parallel_array	mom2	R	None	12:14:27	Tomorr 00:14	10:35:51	100	6044
119315	julian	director903	runh3oco3_460	mom1	R	None	09:26:22	21:26:23	7:47:46	500	9680
119330	nqazi	director893	jetfan_2d	mom2	R	None	08:25:26	20:25:26	6:46:50	1	1942
119316	julian	director903	runh3oco3_460_	mom1	R	None	07:38:13	19:38:14	5:59:37	50	2115
119339	mthatche	director899	ccam768	mom1	R	None	08:38:23	18:38:24	4:59:47	576	30311
119329	jzanotti	director898	bqcd	mom1	R	None	08:24:25	16:24:26	2:45:49	192	10677
119542	pryan	director100	bash	magnus	R	None	13:29:46	14:29:46	51:10	1	1050
119532	charris	director100	bash	magnus	R	None	13:02:11	14:02:11	23:35	1	1034
119544	mshaikh	director100	ddt	mom1	R	None	13:34:27	13:44:28	5:51	1	1034

sacctmgr

- Django user registration / proposal site
- Once approved we create user account on system
- Script calls sacctmgr and sets up association betweeen project account, user and cluster.
- Fairshare=parent
  (PI problem if user burns allocation)

Inter-Cluster Dependencies

- Each Cray (Magnus, Galaxy, Chaos) defined as a slurm cluster
- Additional ‘Zeus’ Cluster of X86_64 IB nodes (contains zythos partition)
- Cray es-DM nodes needed for ‘copyq’ use for users to run serial stage in/stage out jobs
- Added magnusdm pawseydm partitions to ‘pawsey’ misc cluster

- *sbatch –M pawsey –p magnusdm stagein.sbatch*
- How to trigger on successful completion to *sbatch –M magnus –p workq myjob.sbatch*?

ivec.org iVEC Advancing science through supercomputing

(not so) Fairshare

- ACTUAL – 14 day halflife,
  - PriorityWeightAge=1000
  - PriorityWeightFairshare=1000
  - PriorityWeightJobSize=100000
  - PriorityWeightPartition=1000
  - PriorityWeightQOS=0
- IDEAL – User has allocation budget (X CPU h).
  When 0, fairshare at low priority.
- GrpCPUMins stops at 0 – used to ‘terminate’ accounts
- Script to set GrpCPUMins to -1 but alter QoS?
- Using Allocation units as Fairshares enough?

Wishlist

- Fairshare
- FlexLM (hello fluent)
- Native Slurm – Stable? Other user feedback?
  - DDT? FlexLM?
- Cluster Compatibility Mode (CCM)
- Roadmap / Release process
- ‘Pretty’ reporting tools for PIs (with % accurate!)
- Incorporate Power usage from Cray envmon

iVEC Build System

- Bash script used to build <software> <version> with the various compilers and other module dependencies
- Install to /ivec/${IVEC_OS}/{tools,apps,bio-apps} or to /group/${project}/software/
- Generates modulefile
- Generates wiki page to paste onto portal

ivec.org iVEC Advancing science through supercomputing

Any Questions?

ivec.org iVEC Advancing science through supercomputing

```
aelwell@magnus:~$ sreport -t hours cluster Utilization start=2014-09-21
end=2014-09-22
--------------------------------------------------------------------------------
Cluster Utilization 21 Sep 00:00 - 21 Sep 23:59 (86400*cpus secs)
Time reported in CPU Hours
--------------------------------------------------------------------------------
  Cluster    Allocated       Down  PLND Down         Idle  Reserved     Reported
--------- ------------ ---------- ---------- ------------ --------- ------------
   magnus       797841      82511          0       766175     67649      1714176


aelwell@magnus:~$ sinfo -e
PARTITION AVAIL JOB_SIZE  TIMELIMIT   CPUS  S:C:T   NODES STATE      NODELIST
workq*    up    1-1480     12:00:00     48 2:12:2       4 down*     nid00[408-411]
workq*    up    1-1480     12:00:00     48 2:12:2    1465 allocated
nid0[0016-0063,0068-0127,0132-0138,0143-0255,0260-0319,0324-0403,0412-0767,0776-0831,0836-0895,090
0-0926,0928-1023,1028-1072,1075-1087,1092-1535]
workq*    up    1-1480     12:00:00     48 2:12:2      11 idle
nid0[0139-0142,0404-0407,0927,1073-1074]
debugq    up    1-8         1:00:00     48 2:12:2       8 idle      nid000[08-15]
```

ivec.org iVEC Advancing science through supercomputing