Wilson Cluster Roadmap

Jim Simone, Glenn Cooper

March 2, 2023

Wilson Facility

- The Wilson cluster facility is an on-premises facility designed for high performance computing.
- The Wilson cluster (WC) is available to the entire Fermilab community.
- It is used by scientists and engineers from many lab quadrants.
- Procedural barriers to using WC are intended to be low.
- The Wilson cluster offers CPU and GPU parallelism, high bandwidth low-latency InfiniBand networking, a high-performance Lustre filesystem.
- Use cases:
  - Medium-scale (<1000 cores) parallel production workflows
  - Fermilab on-ramp to big HPC centers, good for workflow development
  - Software development and benchmarking, offers both CPU and GPU
  - AI model training – multi-GPU capable for larger models

WC Roadmap issues

- The current WC cluster must soon be moved from GCC/CRC to GCC/CRB
- WC hardware is beyond warranty – need a plan to refresh CPU/GPU hardware
- What is the funding / charge model for WC?

Existing Wilson Cluster Hardware

GCC - CRC					
Function	**Type**	**Name**	**Qty**	**Warranty?**	**Move?**
Login	Server	wc, wc2	2	N	Y
SLURM		wcs, wcce01	2	N	Y
External I/O		wcio	1	N	Y
NFS		wcnfs	1	N	Y
CPU	Worker	wcwn001–100	100	N	Y
CPU	Worker	wcibmpower01	1	Y	Y
CPU	Worker	wcknl01	1	N	N
K40 GPU	Worker	wcgwn001–029	28	N	N
P100 GPU	Worker	wcgpu01–02	2	N	Y?
V100 GPU	Worker	wcgpu03–06	4	Y	Y
A100 GPU	Worker	wcgpua01	1	Y	Y
Lustre		wclfs[m,b,0–3]	6	N	N
Network	Infiniband	Mellanox QDR		N	Y

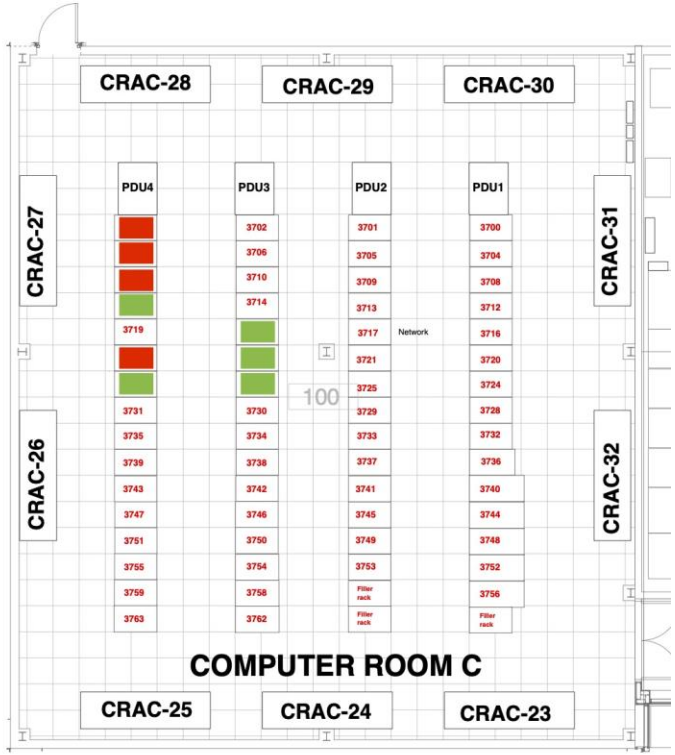
GCC - CRB			
Function	**Name**	**Qty**	**Warranty?**
Lustre	wclfs-new-*	6	Y

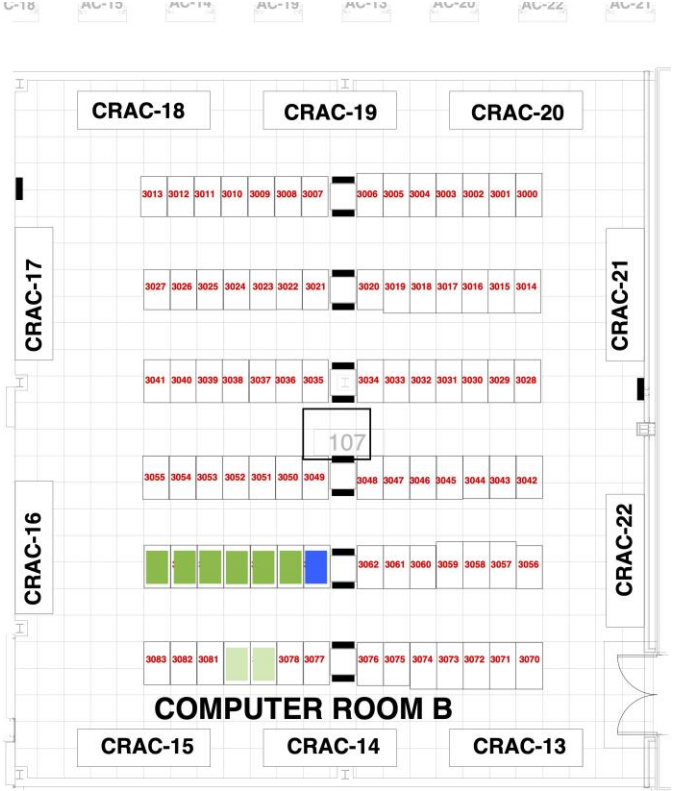
CPU workers dual 8-core Intel "Ivy Bridge" (2012)
NVIDIA A100-80 four / 1 worker
NVIDIA V100 eight / 4 workers
NVIDIA V100 four / 1 IBM Power9 worker
NVIDIA P100 eight+two / 2 servers
NVIDIA K40 obsolete

Summary: 6 servers, 101 CPU, 7 GPU to move

Fermilab

Floor map: Move CRC ➔ CRB





WC Move: Next steps

- Work out details of move: Rack locations, cables, available power, etc.
- We're hoping to contract with Koi Computers, who did the initial installation, to perform the actual move.
- For the move:
  - Bring up new Lustre systems
    - Tell users to migrate any data they want to save by …
  - Choose a date, announce downtime
  - Koi moves everything
  - Bring up elements piece by piece: storage, then head nodes, then workers
  - Perform tests, resolve issues
  - Announce availability
- Data Center Operations can then turn off GCC-CRC.

WC Hardware refresh / funding model

- WC system designed with features needed to run parallel jobs with scalable performance.
- Desired features
  - CPU workers offering good memory bandwidth per CPU core; not just core count / GHz.
  - Low latency high bandwidth network (InfiniBand) essential to scale MPI performance.
  - High I/O bandwidth parallel filesystem.
  - Multiple GPUs / node to scale HPC workloads and AI training.
  - High bandwidth low latency access to GPUs, both inner- and intra-node, and to I/O.
- Want mix of CPU and GPU workers to meet future demands.
- AMD, Intel, and NVIDIA have appealing products integrating CPUs and GPUs.
- How will Fermilab fund new hardware? When? FY23, -24??
- CSAID to develop a formal funding / charge model for WC and other facilities?

Backup Slides

WC Top users / Distribution of job core counts

Top 30 Users 2021-10-01T00:00:00 - 2023-02-27T23:59:59 (44499600 secs)
Usage reported in CPU Hours/Percentage of Total

Account	Proper Name	Used
beamopt	Eremey Valetov	1978623(6.24%)
qcdloop	John Campbell	1871349(5.90%)
qis_algo	Andy C. Y. Li	1377262(4.34%)
ewpht	Tong Ou	914862(2.88%)
grid_pilots	osg	864161(2.72%)
oscisims	Kevin Kelly	864046(2.72%)
qcdloop	Tobias Neumann	836688(2.64%)
sherpa	Stefan Hoeche	805340(2.54%)
desy3	Kenneth Herner	538781(1.70%)
dm_pheno	Tanner Trickle	395714(1.25%)
nuself	Subhajit Ghosh	353486(1.11%)
g4p	Julia Yarba	289913(0.91%)
miniboone	Paul Lebrun	260518(0.82%)
bbind	Michael Wagman	188352(0.59%)
accelsim	Eric Stern	169007(0.53%)
qis_algo	Alessandro Berti	150919(0.48%)
bbind	Benoit Assi	124820(0.39%)
iota	Jean-Francois Ostiguy	103168(0.33%)
qcdloop	R. Keith Ellis	59923(0.19%)
iota	Dmitry Shatilov	47413(0.15%)
fwk	Kaushal Gumpula	38553(0.12%)
genie	Stephen Mrenna	31071(0.10%)
iota	Nilanjan Banerjee	29229(0.09%)
g4p	Soon Yung Jun	28647(0.09%)
g4v	Julia Yarba	28579(0.09%)
fwk	Xinyuan You	24399(0.08%)
wc_test	James Simone	17579(0.06%)
nova	Derek Doyle	17131(0.05%)
fwk	Daniel Grzenda	13145(0.04%)
lqcd_bench	James Simone	11799(0.04%)