Excerpts from: High End Computing File Systems and I/O (HECFSIO): Coordinating the US Government Research Investments

Gary Grider, James Nunez, John Bent, Los Alamos National Lab
Robert Ross, Argonne National Lab
Lee Ward, Sandia National Labs
Steve Poole, Oak Ridge National Lab

HEC Inter Agency Working Group
May19, 2011

State Before HECFSIO/HECURA

- Pre-mid 90's interesting storage research funded by DARPA (RAID, NASD, Active Disk, etc.)
- Mid 90's - 2005
  - Very small, disjoint HEC File Systems and I/O community
  - 3-4 National Labs had tiny teams each, mostly in middleware
  - 3-4 University storage research centers that dabbled in High End Computing needs
  - Mostly individual partnerships with DOE National Labs and ASCI Level 2/3 alliance partners
  - No real organized community with deep centers and broad/diverse draw
  - Problem set addressed mostly addressed by vendors and "Path Forward" like activities
  - The compute to IO gap was beginning it was too early to see how profound of a problem set we faced

The Ever Widening Compute to IO Gap

- Number of processing (cores) units are growing fast
- Memory per core is not getting larger, in fact it will likely get smaller
- Scale up of machines means MTTI will continue to decrease
- Disks are getting faster, but not at same rate as disk density growth, nor at the same rate as Machine/Core/Total Memory growth
- Sweet spot block sizes for disks/raids are getting bigger
- We have to add disk units at a faster pace than we add processor units to stay balanced.
- We have to gather more smaller and less aligned data from more cores and realign/aggregate for more disk devices.
- Maybe this is more like a train wreck than a gap!

Learning From our Failures and Successes

- In the last decade, the IO challenge was formidable and we rose to that challenge with trial and error,
  - using government labs/agencies,
  - vendor partnerships, and
  - university research
- We learned that in absence of an infinite budget, a coordinated effort is essential
- Unfortunately, due to that ever widening gap, the challenges that face us make the last decade look like a cake walk!

HECFSIO is the Framework for Coordination of Efforts in this Vital and Difficult Area of Pursuit

What are the Areas to Address in HECFSIO?

- Global Parallel File Systems/Archives
- Many Terabytes/sec
- 100’s of thousands to millions of metadata ops/sec
- Millions of processes opening/writing/reading
- Millions to Billions of files in a directory
- Trillions of files in file system
- 10’s-100’s of thousands of storage devices
- 10-100 GB/sec archives

Who and What is HECFSIO

- Who
  - DoE: Office of Science and NNSA
  - NASA
  - DoD
  - A few others have dabbled
- To coordinate government investments on File Systems and I/O R&D by the government agencies that are involved in HEC
  - NSF, EPA, DOE: Office of Science & NNSA, DoD, NASA, Others

How Does HECFSIO Succeed?

- The Right Recipe of Ingredients!
  - Experts to guide/coordinate investments/strategy
  - Setting/Vetting/Conveying the nation R&D agenda
  - Forming/cultivating community to attack the national R&D agenda
  - Enable that community for the attack
  - Living/Evolving effort with artifacts and metrics/tracking progress
  - Investments of time/funds until fundamental solution space understood (this has not occurred yet)
    - Realizing that this area is far from having a fundamental solution set, the challenges are getting more formidable at least in the coming decade
    - Realizing that these multi-decadal problem areas require multiple generations of expertise with deliberate knowledge management

Experts

- Hard work by many
  - Rob, ANL
  - Evan, PNNL
  - Lee, Mathew, Sandia
  - Galen, Steve ORNL
  - James, John, Gary LANL
  - Marti, Steve DoD
  - Ellen, NASA

Setting/Vetting /Conveying

- Right people
  - National experts from government/industry/academia
- Right vehicles
  - Designated National Workshop
  - Quality and freely available documentation of National R&D Agenda and related discussions
  - Reach, Interest and Rewards
    - Interesting Problems is easy in this space
    - Targeted enough to convey the National R&D Agenda and funded enough to convey importance
    - NSF's voice to the broad Research Community that HECFSIO is important

Conveying the National Agenda, Gaps, and Overlaps

- Categories of need
  - Metadata
  - Measurement and Understanding
  - Quality of Service
  - Security
  - Next generation I/O architectures
  - Communication and Protocols
  - Management and RAS
  - Archive

- HECFSIO Roadmaps created for quick understanding of research gap area, and importance

Major Gap Area

Current Researchers

Duration of Funding

Comments from Government Agencies

Problem or Area of concentration

2007 QoS Gap Area

Area	Researchers	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
End to End QoS in HEC	Brandt							Good research, but much work needed to get a standards based solution.
	Chiueh							
	Ganger							
Standard API for QoS	SciDAC - PDSI							Very partially addressed by proposed HEC POSIX Extensions. Will be driven by above "End to End QoS in HEC".
	POSIX HPC Extensions							
	PVFS							

Rankings by Governments Agencies

How Important to Government Agencies?

Very Important
Medium Importance
Low Importance

Greatly Needs Research
Needs Research
Does Not Need Research

Greatly Needs Commercialization
Needs Commercialization
Does Not Need Commercialization

Full Calendar Year Funding
Partial Calendar Year Funding
On-Going Work

Is Research Needed in this area?

Is Research ready for commercialization?

Forming/Cultivating a Community to Attack the National R&D Agenda

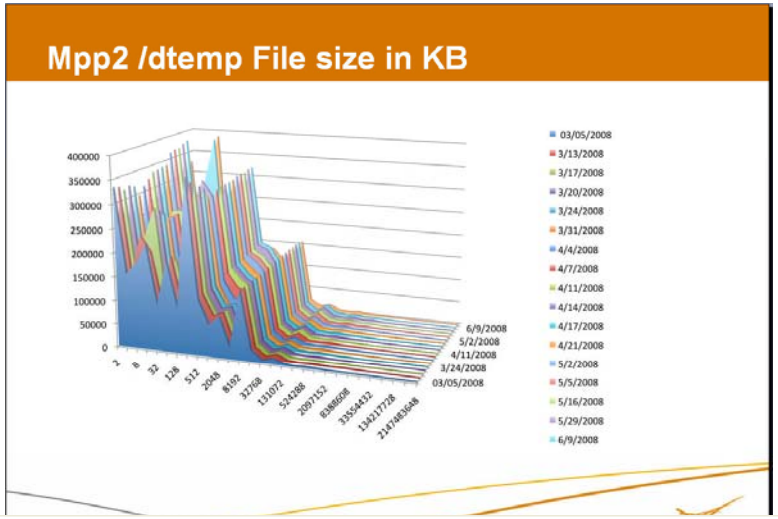
- Reach and Depth
  - DOE/DOD reach into the research community is limited to top few institutes in this area and it is important to have large enduring relationships with these institutions.
  - For every great idea it takes a hundred good ideas…
  - NSF is the key to the larger reach to grow and sustain diversity. This is crucial to the health of the community.
- Collaboration and Understanding
  - National Workshops for Exposure/Collaboration Cultivation
  - Experts involvement, interest, guidance, visits etc.
    - HECFSIO DOE/DOD members tried to visit every HECURA project at least once during the project life (you probably didn't know this)
- Consistent reason to choose this area over other problems to pursue
  - Interesting, evolving and conveyed problem set
  - Funding opportunities to promote interest

Enabling that Community for the Attack

- Resources
  - Funding
    - DOE/DOD for deep enduring
    - NSF for broad and diverse
  - A name, a face, an email address to contact for further understanding of the problem
    - This is something DOE is essential for
  - Experimental Facilities
    - NSF sites like PRObE targeted at systems level CS research
    - DOE Incite to a lesser degree due to domain science focus but still important that CS be one of the many domains
    - DOE/DOD testbeds
  - Data, data, data about the problems and tools to explore the problems

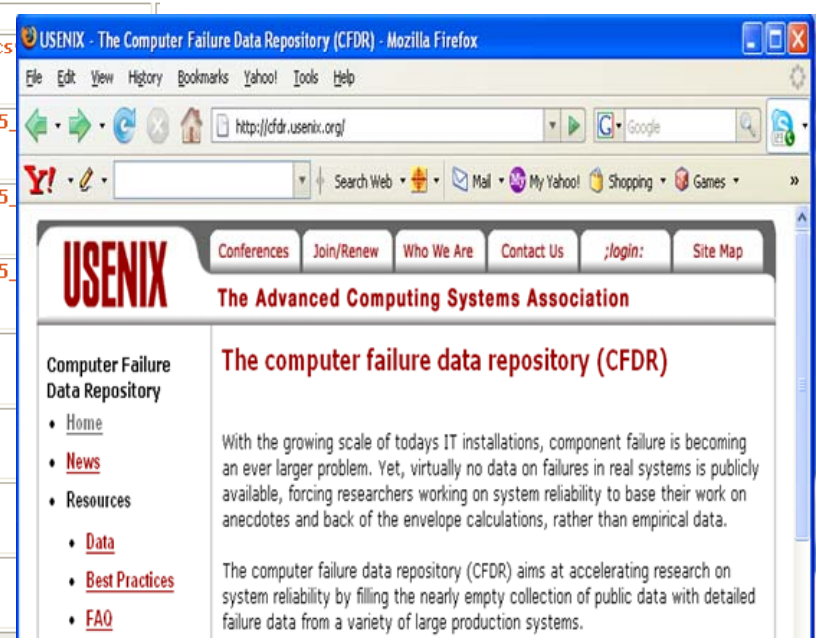
Data, Data, Data and Tools

- **Nothing helps a researcher like real data about the problem!**
- **HECFSIO has promoted this extremely successfully and it has paid off in spades!**



- Failure, usage, placement, file system stats, traces, logs, environmental (heat etc.)
- Updated/refreshed
- Trace mechanisms, visualization tools, simulation tools
- Pseudo applications/kernels
- Broadly available and disseminated

description	size	records	name
all systems failure/interrupt data 1996-2005	2963538	23741	LA-UR-05-7318-failure-data-1996-2005.cs
system 20 usage with domain info	51675641	489376	LA-UR-06-0803-MX20_NODES_0_TO_255_
system 20 usage with node info nodes number from zero	43926669	489376	LA-UR-06-0803-MX20_NODES_0_TO_255_
system 20 event info nodes number from zero	33120015	433490	LA-UR-06-0803-MX20_NODES_0_TO_255_
system 15 usage with node info nodes number from zero	2416139	17823	LA-UR-06-0999-MX15-NODE-Z.TXT
system 16 usage with node info nodes number from one	321293488	1630479	LA-UR-06-1446-MX16-NODE-NOZ.TXT
system 23 usage with node info nodes number from one	60674531	654927	LA-UR-06-1447-MX23-NODE-NOZ.TXT
system 8 usage with node info nodes number from one	67291020	763293	LA-UR-06-3194-MX8-NODE-NOZ.TXT



Living/Evolving effort with Artifacts and Metrics/Tracking Progress

- HECURA solicitation in 2005 received 62 proposals from 80 universities, of which 23 were funded.
- HECURA solicitation in 2009 received 109 proposals from 54 universities, of which 20 were funded.
- HECURA Projects have
  - published approximately 250 conference and journal proceedings
  - contributed 13 open-source software packages
  - Funded over 50 professors and 100 students
  - Have produced 16 Ph.D. and 22 Masters graduates
  - 27 of these graduates have brought their expertise in these problems of national interest to industry,
    - 4 to government
    - 11 to academia – ensuring that future generations of students will be familiar with the area and prepared to attack key challenges.
- At least 3 commercial companies have emerged from these projects, providing valuable and supported products as a return to the community.

- Workshop Attendance
  - 2006 88
  - 2007 99
  - 2008 100
  - 2009 99
  - 2010 114
  - Base set returns, many new faces each year (DOE/DOD enduring/NSF broad and diverse)

Documentation/Dissemination

- Reasonably complete and frequently visited web site
- Religious annual update of
  - Historical report updated
  - Workshop proceedings
  - Roadmaps updated
- HECURA and other projects/abstracts, notes from site visits
- Pointers Data, Data, Data and Tools
- Pointers to Resources
- Standards like NFSv4/pNFS etc.

Summary

- Before HECFSIO we were ill equipped to address this national scale multi-decade problem set
- The gap is just getting wider and wider
- We have a recipe for success for keeping this crucial problem set in check, while we seek fundamental solutions, as long as the ingredients are available!
- This is an example of Inter-Agency teamwork, leverage of each others strengths

- Thank you for all the support you have provided over the years, we hope to continue the good fight!

Discussion: Funding/Emphasis Shifts

- Given it is unclear whether NSF will be supporting a more focused HEC FSIO oriented call which implies funding will be in much more broad calls
- Given ASCR seems to be headed towards more HPC oriented solicitations with winning projects being teams including labs and universities.
- We don’t think that
  - The researchers will pose HEC FSIO Research to broad NSF calls due to popularity/potential reviewer expertise/etc.
  - The broader set of Researchers will have a chance at partnering with labs successfully on ASCR calls
  - The community will continue to attend workshops like this one since there wont be much evidence that there really is HEC coordination needed or wanted

Discussion: Use of HECFSIO Products

- Given it is unclear whether NSF will be supporting a more focused HEC FSIO oriented call which implies funding will be in much more broad calls
- Given ASCR seems to be headed towards more HPC oriented solicitations for the labs (who already understand HECFSIO problems) to compete over
- We question the need for a HECFSIO advisory group and its activities (its not free, it takes time and money from some of the top people in the world)
  - Will any agency ever look at our reports/roadmaps again?
  - If so, for what purpose?
  - If the funding profile changes as in the last slide, will there be a chance of keeping a community, keeping our problems front and center, helping provide data/resources?

Discussion: Continue/Change/Die

- Should we continue this HECFSIO effort?

- Should we change what we do or how we do it?

Thanks

For More Information

HEC FSIO Workshops

http://institutes.lanl.gov/hec-fsio/workshops/

HEC FSIO Roadmaps

http://institutes.lanl.gov/hec-fsio/docs/