File Systems and I/O
A Quick Tour

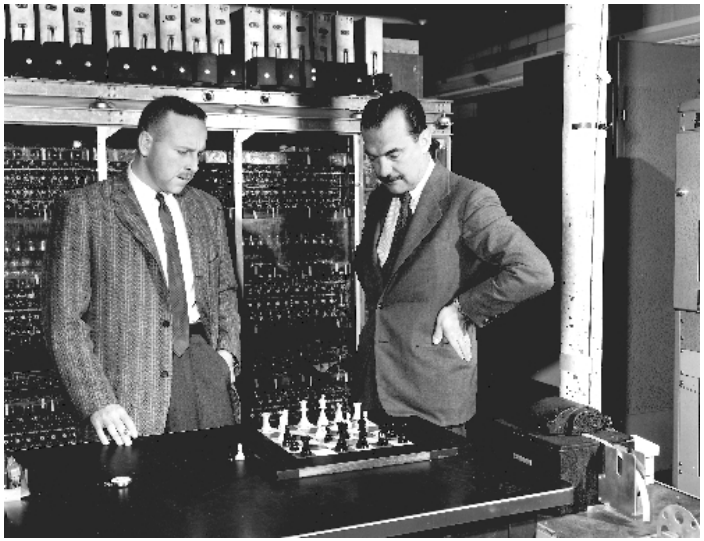
June 10

Gary Grider
Deputy HPC Division Leader
Los Alamos National Laboratory

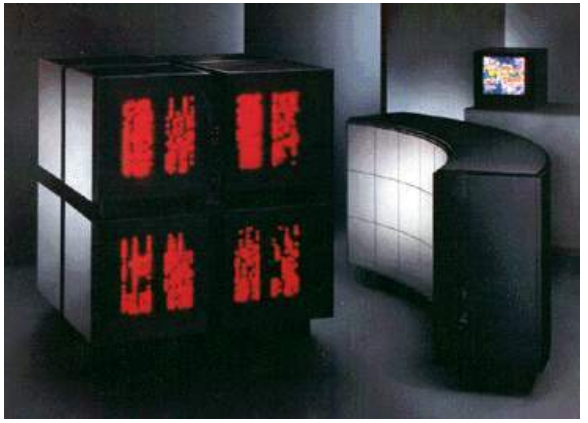
Excerpts from LA-UR-08-2876

Los Alamos
NATIONAL LABORATORY
EST.1943

Kilo, Mega, Gigascale



Maniac1 at Los Alamos
1957 a few tips



1987 CM2 at Los Alamos
64k 1bit procs – 8 gflops
10 GBytes (250 X 40 MB drives) each

1976 Cray 1 at Los Alamos 160 mips

1965 IBM 7030 4 mips
Stretch at Los Alamos

1992 CM-5 25 gflops 1024 sparcs 30 GB (30 X 330 MB drives) each

Terascale

LANL Q – 3TF

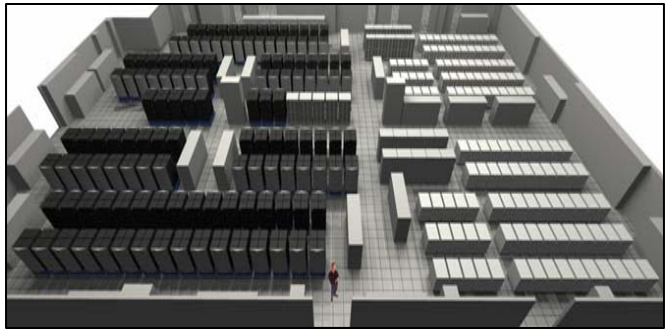
400 TB (5500 X 72 GB drives) 4+1 RAID5

2001

LANL Blue Mountain – 3TF

60 TB (3300 X 18 GB drives) 8+1 RAID3

1996



LLNL White– 7.1TF

20 TB (550 X 36 GB drives) 8+1 RAID3

2000

The Petascale Regime Is Upon Us

The future holds more capability!

	ZIA	TRINITY
Peak PF	> 2	> 50
Total memory	> 0.5 PB	> 5 PB
Aggregate[(a)] Memory BW	> 1 PB/sec	> 5 PB/sec
Aggregate Interconnect BW	> 1 PB/sec	> 7 PB/sec
Aggregate Bisection BW[(b)]	> 80 TB/sec	> 450 TB/sec
Aggregate Message Rate	> 10 GMsgs/sec	> 80 GMsgs/sec
Aggregate I/O BW	> 1 TB/sec	> 10 TB/sec
Disk Capacity	> 20 PB	> 200 PB
System Power (MW)	5 - 8	10 - 16
Floor Space (sq ft)	< 8,000	< 8,000
MTTI (Job) / MTBF (System) (Both @ Full Scale)	> 50 / > 200 Hrs.	> 50 / > 200 Hrs.

Why do I care about HEC FSIO?

- 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 disks
- 10-100 GB/sec archives

The Formation of the HEC FSIO

- The President's Information Technology Advisory Council and White House Office of Science and Technology Policy
- The High End Computing Revitalization Task Force (HECRTF)
  - engage in planning activities to guide future investments
- Interagency Working Group on HEC (HECIWG)

HEC FSIO Current Information

- Categories of needed research:
  - Metadata
  - Measurement and Understanding
  - Quality of Service
  - Security
  - Next generation I/O architectures
  - Communication protocols
  - Management and RAS
  - Archive
- Accomplishments in 3 years:
  - 3 national workshops
  - $15M+ in NSF HECURA and CPA I/O and File Systems research awards - 29 projects
  - $25M SciDAC2 I/O and File Systems related research 5 year awards – 2 projects (SDM Center and PDSI)
  - Simulation resources – Incite and NSF infrastructure
  - $1M DOD ACS I/O – 3 awards
  - Massive amount of failure, usage, event, and parallel trace data released
  - Progress on relevant standards – pNFS and POSIX HECEWG
  - Help Universities with storage, file system, and I/O programs – ISSDM

2007 Metadata Gap Area

Area	Researcher	CY 06	CY 07	CY 08	CY 09	CY 10	CY 11	Rankings
Scaling	Bender							All existing work is evolutionary. What is lacking is revolutionary research; no fundamental solutions proposed.
	Leiserson							
	Maccabe/Schwann							
	SciDAC - PDSI							
	HECEWG HPC Extensions							
	UCSC's Ceph							
	Lustre							
	ANL/CMU – Large Directory							
	PVFS							
Extensibility and Name Spaces	Bender							All existing work is evolutionary.
	Leiserson							
	Tosun							
	Wyckoff							
	UCSC – LiFS/facets							
	ANL/CMU - MDFS							
	SciDAC PDSI							
File System/ Archive Metadata Integration	Lustre HSM							Extended Attributes, although not standardized, could solve problem.
	UMN Lustre Archive							
Hybrid Devices Exploitation	*None*							Research is being done, but no research focused on metdata
Data Transparency and Access Methods	*None*							No research focused on metdata

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

2007 Assisting with Standards, Research and Education

Area	FY07	FY 08	FY 09	FY 10	FY 11
Standards: POSIX HEC ANSI OBSD IETF pNFS	 PDSI U Mich CITI patch pushing/maint Revamp of man pages V2 nearing pub V 4.1 nearing pub Assistance in testing may be needed	 First Linux full patch set Some file system pilot test Initial products			
Community Building	*HEC FSIO 2007* *HEC presence at FAST and IEEE MSST*	*HEC FSIO 2008* *HEC presence at FAST and IEEE MSST*	*HEC FSIO 2009* *HEC presence at FAST and IEEE MSST*	*HEC FSIO 2010* *HEC presence at FAST and IEEE MSST*	*HEC FSIO 2011* *HEC presence at FAST and IEEE MSST*
Equipment	*Incite and NSF Infra* *Need scale CS disruptive facility*	*Incite and NSF Infra* *Need scale CS disruptive facility*	*Incite and NSF Infra* *Need scale CS disruptive facility*	*Incite and NSF Infra* *Need scale CS disruptive facility*	*Incite and NSF Infra* *Need scale CS disruptive facility*
Simulation Tools	*Ligon* *PDSI Felix/Farber*	*Ligon* *PDSI Felix/Farber*	*Ligon* *PDSI Felix/Farber*		
Education	*LANL Institutes as one example* *PDSI*	*Other Institute like activites*			
Research Data	*Failure, usage, event data*	*Many more traces, FSSTATS, more disk failure data*			

Enable Others: A plug for benchmarks, traces, and Kernels!

- Benchmarks
  - Benchmarks suffer from lack of realist representation of real workloads but if you have benchmarks, please consider opening them up to allow others to help
  - Often Benchmarks are “micro-benchmarks” where one portion of a large workflow is represented. We really need to start thinking about capturing workflow which could lead to “macro-benchmarks” which could represent the entire workflow – which is much closer to the bottom line
- Traces
  - Can be done in such a way that gives a reasonable representation of your workloads (except in the highly parallel world this is harder). This can be very helpful for people to help you. Can be anonomized.
  - There are clearinghouses for traces.
- Kernels
  - Takes work to provide but can really help others help you
  - There are clearinghouses that would be happy to take these as well

Resources

- HEC FSIO planning site
  - http://institute.lanl.gov/hec-fsio/
- ISSDM site
  - http://institute.lanl.gov/isti/issdm
- PDSI site
  - http://institute.lanl.gov/pdsi