



Integrated Computing & Communications
Scalable I/O Project

Scalable I/O Project Overview and FLASH I/O

This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.



Integrated Computing & Communications
Scalable I/O Project

Today's Agenda

Introduce the ASCI Scalable I/O Project

Discuss I/O performance and usability

Discuss the I/O model in the FLASH code

Invite you to involve us in your project



Integrated Computing & Communications

Scalable I/O Project

Organization

Computation

- ▣ Integrated Computing and Communications
 - ▣ Services and Development Division
 - ▣ Development Environment Group
 - ▣ **Scalable I/O Project**



Integrated Computing & Communications

Scalable I/O Project

Activities and Responsibilities

Procurement

- ▾ Evaluation and acceptance testing
- ▾ Ongoing characterization of performance and stability

Support

- ▾ Exception testing (identify and resolve bugs, test fixes)
- ▾ Products:
 - ▾ GPFS, HDF5, MPI-IO, MPI-IO/HPSS
- ▾ Support customers directly

Development

- ▾ Develop generally applicable high performance I/O solutions (i.e., ASCI I/O stack)

Research

- ▾ Future Parallel File Systems (CFS - LINUX)



Integrated Computing & Communications
Scalable I/O Project

Team Member's Specialties

Tyce McLarty (Project Leader)

- ▾ HDF5, SAF, GPFS, CFS

Richard Hedges

- ▾ MPI-IO, GPFS, CFS, HDF5

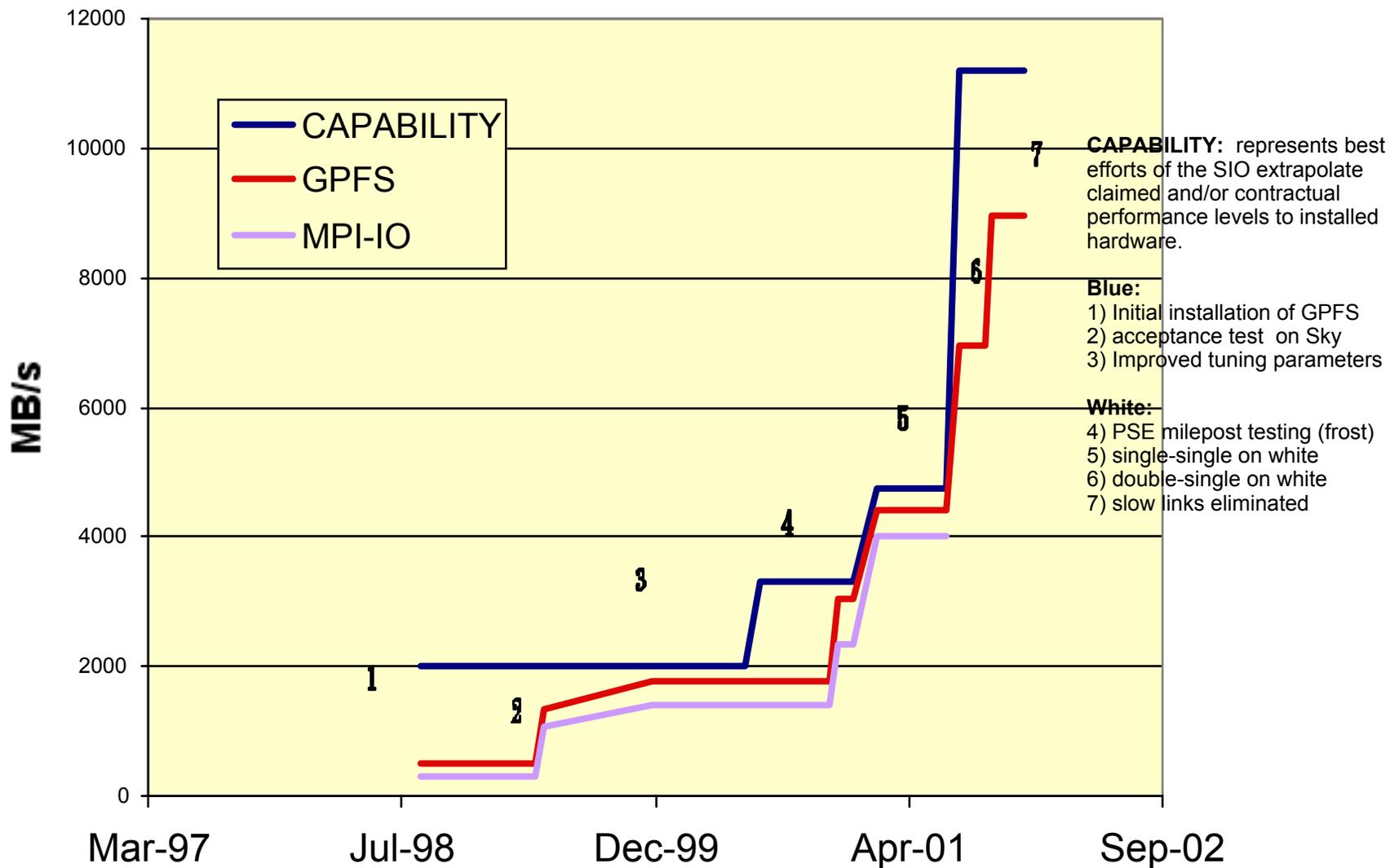
Bill Loewe

- ▾ GPFS, test development, CFS

Linda Stanberry

- ▾ MPI, MPI-IO, MPI-IO/HPSS development

GPFS Progress Timeline





Integrated Computing & Communications

Scalable I/O Project

I/O Issues to Consider

System Issues (peak performance)

- I/O system sized so 1/2 of memory can be dumped in 5 minutes
- TFLOP / GB/sec ratio = 1

Performance (time to solution)

- If I/O is < 10% wall clock time: performance not a problem
- If I/O > 25% wall clock time: possible performance problem
- If I/O > 50% wall clock time: definite performance problem

File System Issues (potential bottlenecks == serialization)

- Many open files - metadata ops
- Single file - fine granularity locking

Usability (workflow for science)

- Ease of restart
- Ease of changing number of processes in simulation
- Ease of post-processing



Integrated Computing & Communications

Scalable I/O Project

ASCI Alliance FLASH Center Coming to Lab

- University of Chicago ASCI Alliance, the “Flash Center”: one of the five ASCI ‘Level-1’ academic alliances, focused on computational grand challenges
-
-
-
- Computer Science: MPI, architectures and parallel computing diagnostics, numerical methods and visualization

and parallel I/O



Integrated Computing & Communications

Scalable I/O Project

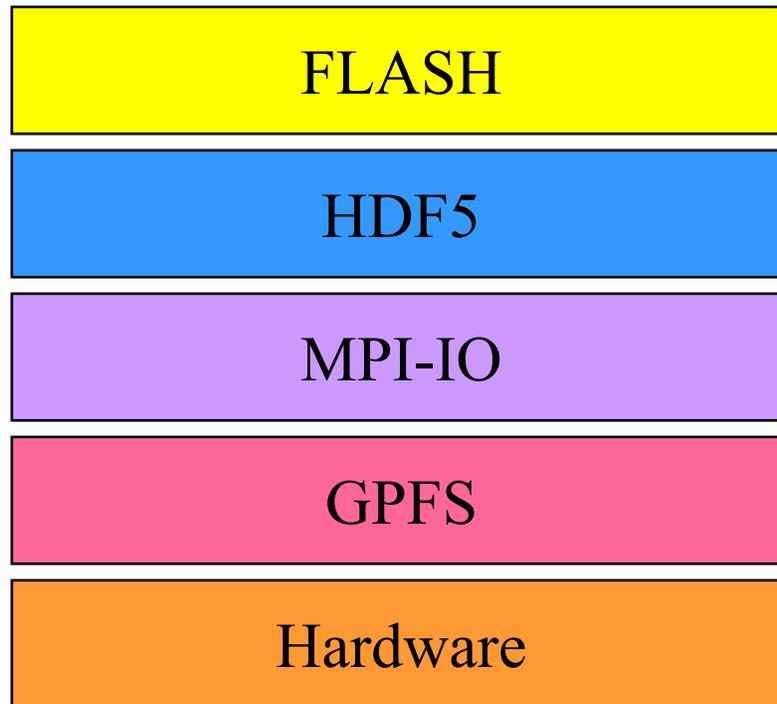
Flash I/O

- ▼ The “typical” FLASH large production run will generate:
 - ▼ .5 Tbytes of data
 - ▼ 1000 plot files
 - ▼ 100 restart files
- ▼ Early runs took as much as 1/2 of the wall clock time for output!
- ▼ FLASH I/O benchmark created to model I/O precisely as in code



Integrated Computing & Communications
Scalable I/O Project

ASCI End-to-End I/O Stack for FLASH





Integrated Computing & Communications

Scalable I/O Project

Solutions

- Wait for complete IBM implementation of MPI-IO (June 2001 on Blue)
- Bug in memory consumption for data type representation in IBM MPI (had been previously identified by SIOP)
- Buffering and alignment of data broken (worked in earlier releases) in HDF5



Integrated Computing & Communications
Scalable I/O Project

Success

- ▾ ASCI I/O Stack Design Implemented in FLASH
- ▾ For 1 restart file, 20 plot files:
I/O time < 5 minutes



Integrated Computing & Communications
Scalable I/O Project

Richard Hedges

richard-hedges@llnl.gov

(925) 423-2699

<http://www.llnl.gov/icc/lc/siop/>