Scalable Systems Software Project

Presented by

Al Geist

Computer Science Research Group
Computer Science and Mathematics Division

Research supported by the Department of Energy's Office of Science
Office of Advanced Scientific Computing Research

SC07

OAK RIDGE National Laboratory

The team
Coordinator: Al Geist

Scalable Systems Software

Participating organizations

NSF Supercomputer Centers

- NCSA
- Pittsburgh Supercomputing Center

DOE Laboratories

- Oak Ridge National Laboratory
- Sandia National Laboratories
- Argonne National Laboratory
- Los Alamos National Laboratory
- Berkeley Lab
- Ames Laboratory
- Pacific Northwest National Laboratory

Vendors

- IBM
- Cray
- intel
- sgi

Open to all, like MPI forum

www.scidac.org/ScalableSystems

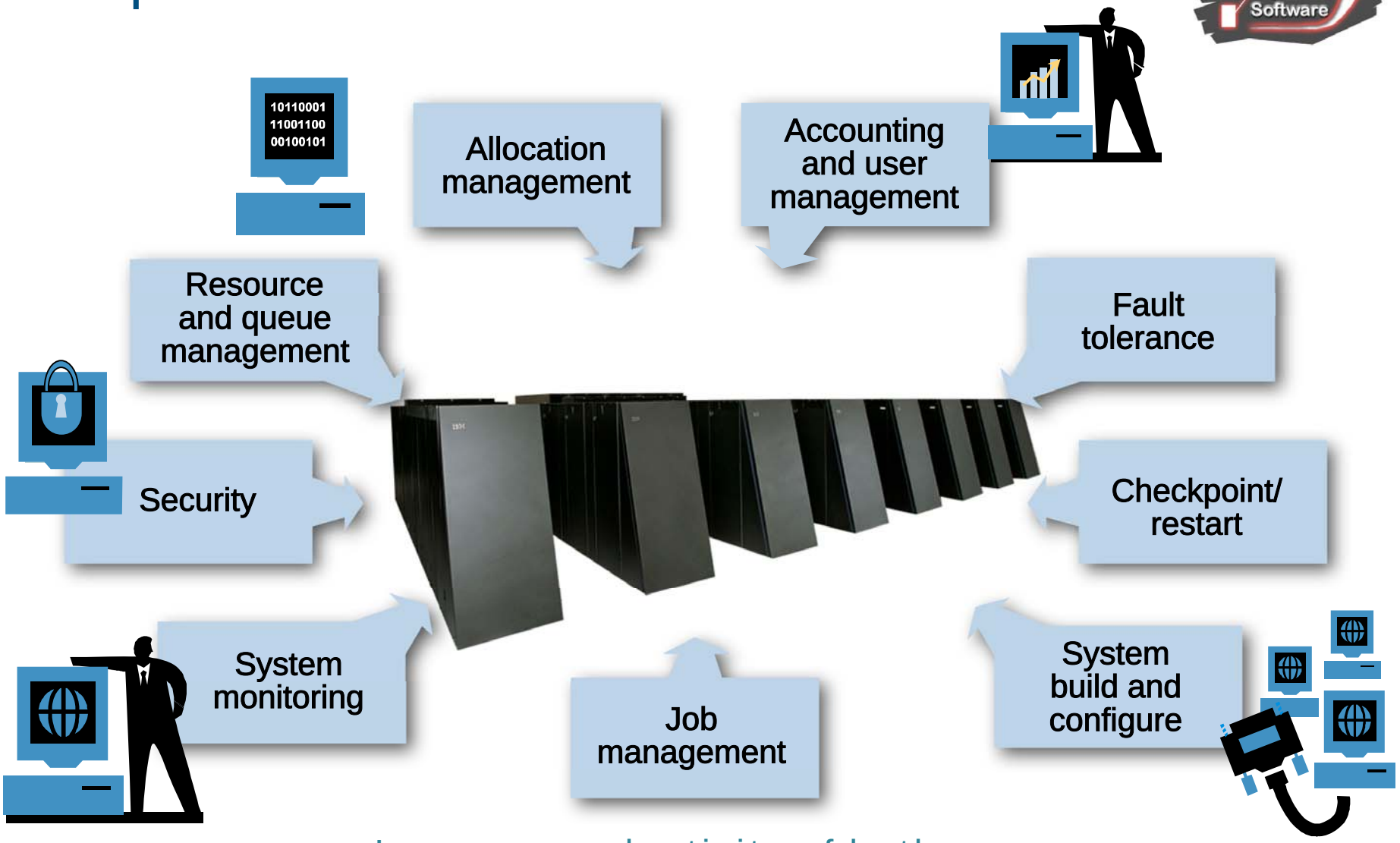
Oak Ridge National Laboratory

The problem

System administrators and managers of terascale computer centers are facing a crisis:

- Computer centers use incompatible, ad hoc sets of systems tools.
- Present tools are not designed to scale to multiteraflop systems – tools must be rewritten.
- Commercial solutions are not happening because business forces drive industry toward servers, not high-performance computing.

Scope of the effort



Improve productivity of both users and system administrators

Impact

Reduced facility management costs

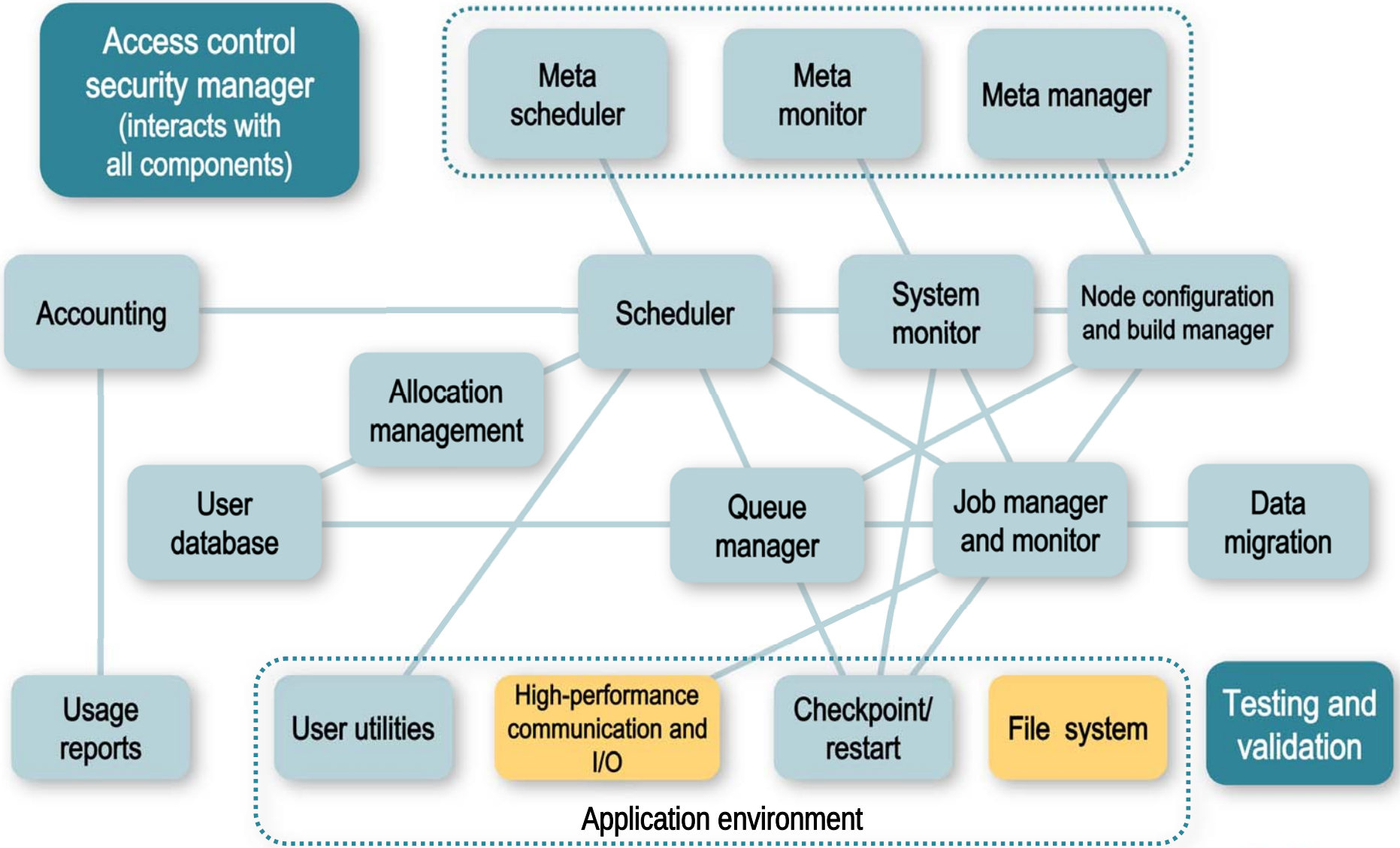
- Reduce duplication of effort in rewriting components
- Reduce need to support ad hoc software
- Better systems tools available
- Able to get machines up and running faster and keep running
- Especially important for LCF

More effective use of machines by scientific applications

- Scalable launch of jobs and checkpoint/restart
- Job monitoring and management tools
- Allocation management interface

Fundamentally change the way future high-end systems software is developed and distributed

System software architecture



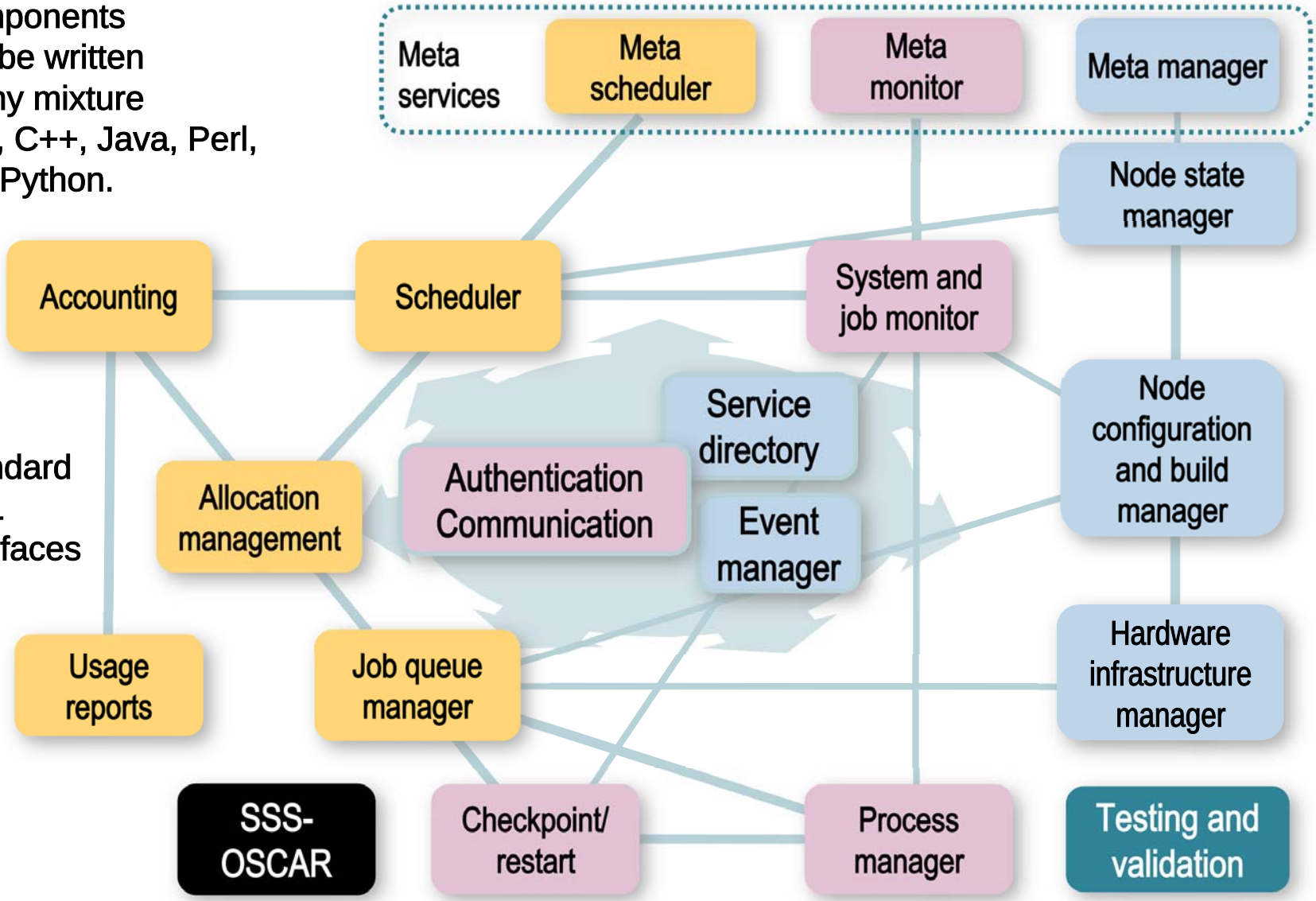
Highlights

Designed modular architecture	• Allows site to plug and play what it needs
Defined XML interfaces	• Independent of language and wire protocol
Reference implementation released	• Version 1.0 released at SC2005
Production users	• ANL, Ames, PNNL, NCSA
Adoption of API	• Maui (3000 downloads/month) • Moab (Amazon.com, Ford, …)

Progress on integrated suite

Components can be written in any mixture of C, C++, Java, Perl, and Python.

Standard XML interfaces



Production users

Running a full suite in production for over a year	• Argonne National Laboratory: 200-node Chiba City and BG/L • Ames Laboratory
Running one or more components in production	• Pacific Northwest National Laboratory: 11.4-TF cluster + others • NCSA
Running a full suite on development systems	• Most participants
Discussions with DOD-HPCMP sites	• Use of our scheduler and accounting components

Adoption of API

Maui scheduler now uses our API in client and server	• 3,000 downloads/month. • 75 of the top 100 supercomputers in the top 500.
Commercial Moab scheduler uses our API	• Amazon.com, Boeing, Ford, Dow Chemical, Lockheed Martin, more…
New capabilities added to schedulers due to API	• Fairness, higher system utilization, improved response time.
Discussion with Cray: Leadership-class computers	• Don Mason attended our meetings • Plan to use XML messages to connect their system components. • Exchanged info on XML format, API test software, more … use of our scheduler and accounting components.

Contact

Al Geist

Computer Science Research Group
Computer Science and Mathematics Division
(865) 574-3153
gst@ornl.gov

OAK RIDGE National Laboratory