

BLUE WATERS

SUSTAINED PETASCALE COMPUTING

VISIONARY COMPUTING



CUG 2018 STOCKHOLM

Best Practices for Management and Operation of Large HPC Installations

Celso Mendes, Scott Lathrop, Jeremy Enos, Brett Bode,
Gregory Bauer, Roberto Sisneros, William Kramer

NCSA



GREAT LAKES CONSORTIUM
FOR PETASCALE COMPUTATION

CRAY

Introduction

HPC centers strive to provide good resources and services

- Involves how to best manage and operate HPC facilities
- Extends to how to develop communities, user support, etc

However...

- These efforts are rarely reported in the literature
- As a result, there is lack of info on good examples to be followed

Goals of this paper:

- Document the best practices adopted with Blue Waters
- Disseminate set of best practices across the HPC community

Introduction (cont.)

Some best practices are directly visible to users

- Portal – allocations and usage, documentation, training
- Monthly calls, webinars, annual Symposia, science highlights
- Points of Contact (POCs) provide direct user support

Other best practices are behind the scenes in support of users

- Strong vendor support for system reliability and availability
- Topology Aware Scheduler
- Security of equipment, intellectual property, and personnel
- Project management

Introduction (cont.)

Blue Waters system

- Largest machine constructed by Cray
- Challenging system to deploy and operate
- Extremely balanced system: computational capability, storage capacity and network connectivity are all very good
 - Allowed effective usage by teams from different scientific areas

Implications of the adoption of best practices:

- Five years of smooth Blue Waters operation so far
- Several documented scientific discoveries by science teams
- Many of these best practices are applicable to other systems too

Areas of Best Practices

Scope: Entire Blue Waters project

- Project Management
- Deployment, Operations and System Management
- Models of User Support
- Communicating Success
- Expanding Current Communities

Best practices adopted in all areas – critical for project's success

In summary: efforts go much beyond machine itself

Format for Documentation of Best Practices

Choice: Quadrant Description

- Simple to understand and use
- Allows more or less detail to be included
- Common across all project areas
- Two sheets created for each Best Practice
 1. Basic information: what it is, relevance, etc.
 2. Salient features of the Best Practice

RELEVANT EXAMPLES OF BEST PRACTICES

(10 examples selected for this presentation – more in the paper)

Best Practice: Enhanced Python Provisioning

What is the Best Practice?

Providing an **expansive Python suite** of many popular software components.

Why is it needed?

Reduces the level of effort by staff and users, improves user experience, and improves performance.

Who does it impact and when?

Users benefit by **spending less time attempting to install python packages** and use **optimally configured packages**.

The center benefits by reducing the number of service requests.

Why is this a Best Practice?

This practice improves the effectiveness of the teams on Blue Waters and allows for complete workflows to be run on the system.

Best Practice: Enhanced Python Provisioning

How is this Best Practice different from the common practice?

- Typical Python installations are composed of base packages plus the “top 6” packages such as: numpy, scipy, cython, mpi4py, matplotlib, and pycuda.

Is it measurable, and if so, how?

- Monitor number of Python service requests (tickets) compared to the number of packages available.

Example of use

- Over 300 Python packages are available, including mpi4py, pycuda, h5py, netcdf4-python, and pandas. The core **numerical libraries** are linked against Intel’s MKL, and the stack is built for running on both login and compute nodes.

Python tickets reduced significantly.

- Anaconda.com provides similar package breadth and flexibility.

Best Practice: User Impersonation

What is the Best Practice?

Members of support staff are **able to impersonate users** on Blue Waters and on the Portal. **Logs are kept for accountability** A **separate shell history** avoids polluting user history.

Why is it needed?

- Can expedite user support when users don't understand how to open directory permissions
- Can allow support staff to build and/or run code in a user's environment, reducing possible degrees of freedom

Who does it impact and when?

- Users requiring assistance building or running code, upon request
- PIs who are temporarily unable to access the Blue Waters Portal

Why is this a Best Practice?

Often, support staff can figure out how to solve the user's help request with much **less delay and email traffic**. It **avoids non-sysadmin staff having root access**.

Best Practice: User Impersonation

How is this Best Practice different from the common practice?

- We are **unaware of the policy** of support staff impersonating users at other sites.

Example of use

- A user is having trouble building a package on Blue Waters, but a staff member can build it with no difficulty in his/her own directory. Impersonating the user, the support staff can locate missing or incorrect modules loaded, interfering with the package build

Is it measurable?

- Each instance of support staff impersonating a user can be logged, in case future review is necessary.

Best Practice: Topology Aware Scheduling

What is the Best Practice?

Recognizing the need for, designing, and utilizing a **network topology aware job scheduler**.

Why is it needed?

It keeps **communication-intensive** applications from overlapping traffic on network paths unnecessarily.

Who does it impact and when?

All users are impacted. Average performance is significantly improved so **system throughput increases**, and runtime consistency rivals that of dedicated system performance.

Why is this a Best Practice?

Without it, the system would have measured fine against the status quo. Both the development and throughput measurement efforts were extensive without a guaranteed result. This venture was a risk, but with an enormous payoff.

Best Practice: Topology Aware Scheduling

How is this Best Practice different from the common practice?

- System utilization emphasis is placed on science throughput instead of node occupancy.

Is it measurable, and if so, how?

- Comparing application runtimes, evaluating system aggregate metrics over time, particularly of the network resource. Value estimate of **increased science delivered** is several million dollars.

Example of use

- Blue Waters compared two 6 month periods and observed a 42% increase in bytes transmitted

Best Practice: Security Model

What is the Best Practice?

Multi-factor authentication (MFA), hierarchical one-way privilege model, centralized control of access and users, and **use of Bro NIDS**.

Who does it impact and when?

Best case: System staff and users alike both stand to lose large amounts of time.

Worst case: Irreplaceable data or IP is lost.

Why is it needed?

Cyber **security threats** and new vulnerabilities are ever present.

Why is this a Best Practice?

MFA is gaining adoption, but represents additional cost and complexity, so it is not universal. MFA use wipes out entire categories of credential theft attack vectors. The one-way privilege model helps ensure isolation should a privilege escalation ever occur.

Best Practice: Security Model

How is this Best Practice different from the common practice?

- MFA is not used everywhere
- **Staff privilege escalation cannot be performed from user accessible points**
- Bro NIDS is not used everywhere. Optical taps **send a copy of all traffic to Bro system** for analysis.

Is it measurable, and if so, how?

- Some types of attacks can be counted. E.g. On average this year, **over 1M IP addresses automatically null routed (blocked)** by NIDS per month. Compromise incidents remain at **zero**.

Best Practice: Heterogeneity

What is the Best Practice?

Providing access to CPU-only nodes, CPU-GPU nodes, or **a mix of both nodes** on a common interconnect and file system.

Why is it needed?

Porting to CPU-GPU architecture while doing production runs on CPU-only nodes improves effectiveness of teams. Allows for CPU-only and CPU-GPU load balance experimentation.

Who does it impact and when?

Projects and users that want to experiment with CPU-GPU architectures.

Why is this a Best Practice?

Improved user experience.
Avoids idling GPU when workloads only run on CPUs.

Best Practice: Heterogeneity

How is this Best Practice different from the common practice ?

- Systems are **typically homogenous** to simplify management of resources.

Is it measurable, and if so, how? →

Example of use:

- Systems such as JUWELS (Julich)

Usage	Sept-Nov 2014	Sept-Nov 2015	June 2015-May 2016
Active Science Teams	79	111	178
Teams running >1 XK job	26	37	73 - (41%)
Teams using > 5 node*hours	23	32	69 - (39%)
Active PRAC Projects	25	35	41
PRAC Teams running >1 XK job	10	12	23 - (56%)
PRAC Teams using > 5 node*hrs	9	12	23 - (59%)
PRAC Teams running > 4,125 XK jobs (average)			5 - (12%)
Major PRAC Teams using > 31,177 XK node-hrs (average)			6 - (15%)

Best Practice: Use of Monitoring

What is the Best Practice?

Provide interfaces to **quickly dive into system monitoring data** correlating compute and storage metric data with job data to easily identify unusual behavior.

Why is it needed?

Blue Waters produces **over 20 billion data points per day**. Gaining insights into this mountain of data is very challenging, but can provide great information system operation and application behavior.

Who does it impact and when?

Staff are able to understand system behavior and diagnose live problems.

Why is this a Best Practice?

Making use of monitoring data is needed to accurately diagnose system issues, reducing interference between users and helping to optimize overall system performance.

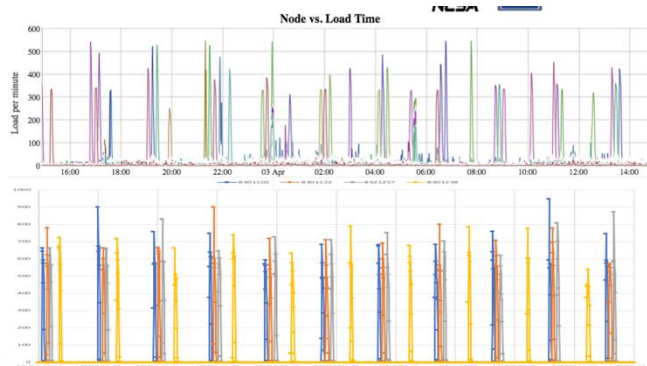
Best Practice: Use of Monitoring

How is this Best Practice different from the common practice?

- Many sites store only summaries of job metrics or only the current live statistics.
- Blue Waters **stores multiple days of full data** in fast storage enabling queries in real time for recent and running jobs.

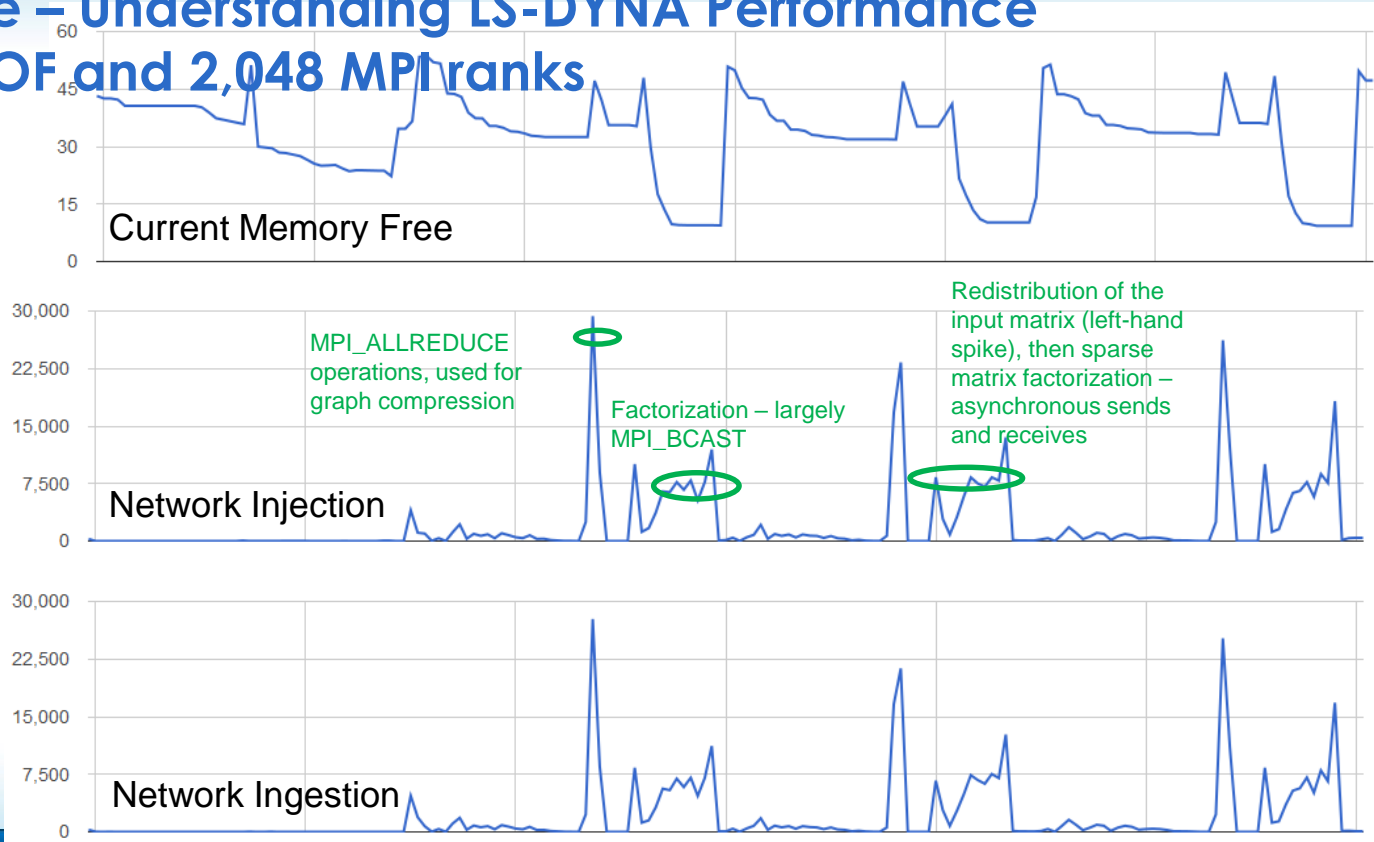
Examples of use:

- Mapping load on Lustre OSTs to IO activity on compute node for a specific job
- Using free memory and network ingestion data to **understand application behavior** (next slide) – No app instrumentation needed!



Example – Understanding LS-DYNA Performance 200M DOF and 2,048 MPI ranks

The spike is MPI_ALLREDUCE
The redistribution spike is asynchronous sends/receives (we also have an MPI_ALLTOALLV variant).
The factorization “ramp up” is largely MPI_BCAST.



Best Practice: Vendor Bug Tracking

What is the Best Practice?

All bugs filed with a vendor have a bug entered into the local bug tracking system. A (semi) automated process **updates local tickets with vendors updates.**

Why is it needed?

Vendors often close bugs as soon as the bug is resolved. The local bug entry allows continued tracking until the update is installed and confirmed. Allows for **broader visibility of Cray issue changes** within local ticket system

Who does it impact and when?

The system ensures that bugs filed by both users and staff are addressed before being closed.

Why is this a Best Practice?

It ensures the fix is available to users and corrects the problem before the bug report is closed.

Best Practice: Vendor Bug Tracking

How is this Best Practice different from the common practice?

- Some other sites create local tickets to shadow vendor tickets, but we are **not aware of any other automated connections**.
- NCSA has an automated connection in place to **update local JIRA tickets with changes in the Cray bug system**.
 - Prior to the change in the Cray system this was done via the nightly change digest that was offered to each user.
 - Now this is done with help from our local Cray site staff via **two email forwardings**.
 - Hint: This could be easier Cray folks!

Best Practice: Risk Register

What is the Best Practice?

The project **tracks all risks** associated with the project. The risks are tagged relative to **the impact on the project, and the probability of occurrence.**

Why is it needed?

Every project has potential risks. The project team needs to be able to monitor all potential risks, and have a plan ready to be implemented to mitigate any risks that are triggered.

Who does it impact and when?

Blue Waters stakeholders, including partners (PI, postdocs, grad students, fellows, interns), staff, NSF, the University of Illinois, and NCSA may be impacted if a risk is triggered and unable to be addressed in a timely manner. Risks are monitored on a continuous basis, and regular quarterly reviews are conducted to adjust as needed.

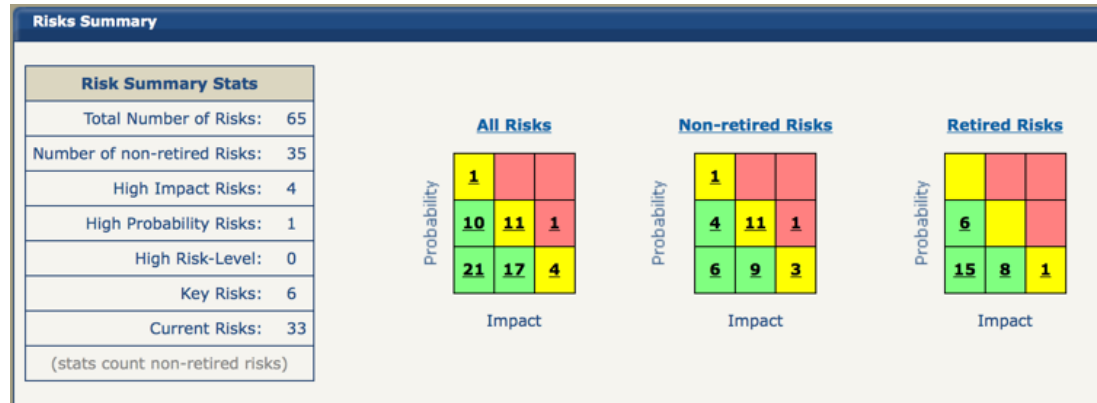
Why is this a Best Practice?

It provides a broad overview of the full range of potential risks for the project, a **trigger to initiate mitigation**, a **plan to mitigate each risk**, a process to adjust the likelihood of each risk, and **process to retire risks that are no longer viable.**

Best Practice: Risk Register

- The following is a screenshot of the Blue Waters' risk tool. When the project first began, there were risks with high probability and high impact.

The process of **tracking the risks** and acting on triggers helped to prevent any serious impacts to the project.



Freely available from: <https://wiki.ncsa.illinois.edu/display/ITS/NCSA+Risk+Register>

Best Practice: Procedure for Acceptance Testing

What is the Best Practice?

General approach for acceptance of new components in Blue Waters or upgrades to existing components, with **detailed acceptance planning and multiple levels of coordination and approval.**

Why is it needed?

To maximize the likelihood that new components will behave as expected and that upgrades will not degrade observed quality of service.

Who does it impact and when?

It can potentially impact every Blue Waters user, especially when there is a major upgrade to an existing component that has been in operation for some time.

Why is this a Best Practice?

It **minimizes the chance for users to be affected** by unexpected bugs in new products, as those bugs would be proactively detected by NCSA staff in early tests.

Best Practice: Procedure for Acceptance Testing

How is this Best Practice different from the common practice?

- Both the design used for a given test (in the acceptance process) and its observed results are evaluated and eventually **approved by multiple specialists** with different perspectives, rather than being restricted to the view of a single judging person.


Example of use

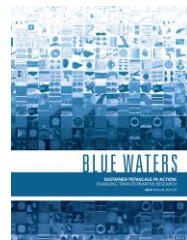
- Acceptance of cabinet expansion in Blue Waters to upgrade its physical topology

Is it measurable?

- Yes – For each test, **assign ranges to the possible results**, then assess overall result by combining individual results obtained according to their ranges.

PLANS FOR DISSEMINATION

- Tutorial presentations – e.g. **PEARC18** 
- Presentations at conferences (e.g. CUG-2018)
- Blue Waters Annual Report/Book: 2-page project summaries
- Blue Waters Symposium (next slide)



Best Practice: Annual Symposium

What is the Best Practice?

Blue Waters hosts an **annual community-driven forum** to discuss petascale and extreme scale challenges and opportunities. **All of the PIs (and their teams) conducting research during the year are invited to provide papers on their research.** Students are encouraged to attend.

Why is it needed?

This provides a unique opportunity for the Blue Waters staff to **meet with the majority of PIs and their research teams** to identify challenges and recommend solutions. The Blue Waters team is able to gather advice and suggestions for improving services, and to gather science stories for the annual report.

Who does it impact and when?

The researchers benefit from learning about cross-cutting challenges and solutions, NSF program officers are able to learn about the computational science being accomplished, and the Blue Waters team gathers insights and advice for future initiatives.

Why is this a Best Practice?

The symposium brings together PIs and their teams from different domains and **fosters scientific and computing collaborations** in a collegial setting where attendees are engaged and focused on the event. Other events are typically domain specific, so the variety of fields of science, with a common set of interests, put BW Symposium apart.

Best Practice: Annual Symposium

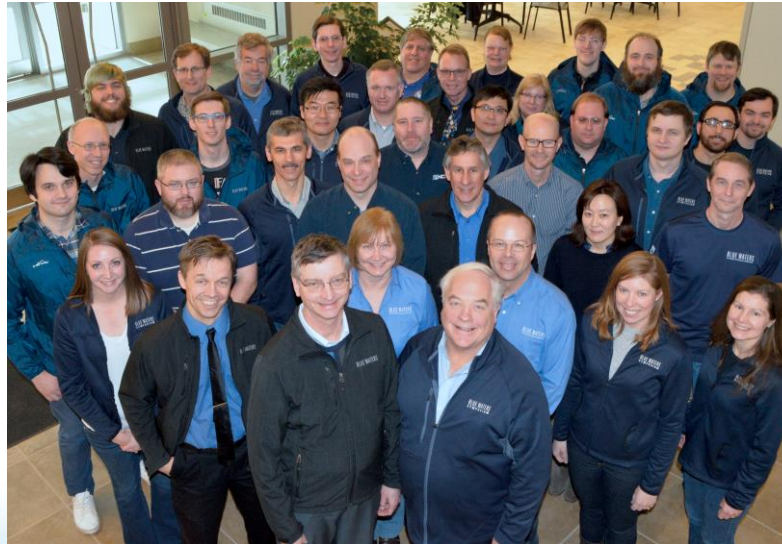
- Over 150 people attend annually, of which over 50 are PIs
- Keynote speakers address “big picture” issues and directions
- Provides a **unique opportunity** to identify petascale and extreme scale requirements, and recommend future directions
- Identifies **improvements** to resources and services
- Fosters the exchange of challenges, opportunities and solutions across **diverse fields of research**

Conclusion

- **Blue Waters Project**
 - Blue Waters System: first sustained-petascale machine, largest by Cray
 - Many other activities – user support, community engagement, vendors, ...
- **Several challenges faced in deployment and operation**
 - Best practices were adopted to allow smooth operation across 5 years
 - Reporting these best practices is not common
- **Our goals in this paper:**
 - Document our experiences, share them with the community
 - Provide guidance to future deployments

Acknowledgments

- Funding: NSF OCI-0725070/ACI-1238993, State of Illinois
- Personnel: **NCSA** Blue Waters team, Cray site team



For Further Information

- Blue Waters Project
 - <http://www.ncsa.illinois.edu/enabling/bluewater>
- Blue Waters Portal
 - <https://bluewater.ncsa.illinois.edu>
- NCSA
 - <http://www.ncsa.illinois.edu>
- Celso Mendes
 - cmendes@illinois.edu