Pacific Rim Application and Grid Middleware Assembly (PRAGMA)\textsuperscript{1}:

PRAGMA Overview, Software, Virtualization and Resource sharing

Philip Papadopoulos, Ph.D.
Chief Technology Officer, San Diego Supercomputer Center
Research Scientist, Calit2
University of California San Diego

\textsuperscript{1} US Participation funded by NSF Award OCI-1234983
Agenda

• Introduction to PRAGMA as an organization
• Scientific Expeditions
  – Lake Ecology
  – Biodiversity
  – Experimental Networking Testbed (Not discussed today)
• Technical Developments
  – Virtual Machines ➔ Virtual Cluster
  – Overlay Networks
  – System Software Packaging
  – Better/easier control of distributed resources
Global Investments and Nature of Collaboration
Why are these important?

• Global problems require global, collaborative responses: investment and collaborations
• Working across disciplines, cultures requires new skill sets
• Working in the *marketplace of ideas* improves quality and helps transfer knowledge more rapidly

We need to articulate the “value proposition” of collaboration
Trends in Collaboration

• The scientific world is becoming increasingly interconnected, with international collaboration on the rise.

• Science in 2011 is increasingly global, occurring in more and more places than ever before.

“intellectual power [is] becoming increasingly evenly distributed”
N. Birdsall, F Fukuyama
Future Human Capital
Why do we need to worry?

• Solving problems, developing technology, making investments and collaborating are done by people

• Are we creating the next generation to face the global challenges, and work in multidisciplinary, multicultural teams?
Environmental and Earth Sciences Data have Characteristics of a Long Tail

- **THE HEAD:**
  - Climate,
  - High Energy Physics

- **THE LONG TAIL:**
  - Environmental and Earth Sciences Data

**Long Tail Characteristics**
- More specialized
- Hard to find
- Lower volume
- Heterogeneous
- Collected by large number of people

Wyborn 2013
Community of Practice
Scientific Expeditions and Infrastructure Experiments
for Pacific Rim Institutions and Researchers

Established in 2002
Welcome

The Pacific Rim Application and Grid Middleware Assembly (PRAGMA) is a community of practice comprising individuals and institutions from around the Pacific Rim that actively collaborate with, and enable, small- to medium-sized groups to solve their problems with information technology. Key to PRAGMA’s success is the active involvement of participants in scientific expeditions, technology development, student engagement, and outreach to new communities.

Attending our workshops is an ideal way to become familiar with what PRAGMA does, and how we interact as a community of trust. We invite you, your students and colleagues to explore this site, visit our partner sites, and attend future workshops (which do not carry a registration fee). We look forward to seeing you there!
Broad Organization

- Scientific Expeditions (focal activity)
- Cooperative technology development
  - Open Source
- Student Engagement
  - Student’s Association
  - Undergraduate summer exchanges via PRIME (US), MURPA/QURPA (Australia)
  - Graduate student visits
- International testbed resources
  - Applications
  - Systems/integration Software
- Steering Committee

Began in 2002 as a small workshop series to better understand “grid” technologies.
PRAGMA Involves People

• Twice a year workshops
• Working groups
  – Resources
  – GEO Sciences and Telescience (Disaster Mitigation)
  – Life Sciences
• Expeditions
  – Lake Eutrophication
  – Biodiversity

• Future Meetings
  – PRAGMA 27, Bloomington Indiana October 15 – 17, 2014
  – PRAGMA 28 Nara Japan April 8-10 2015
PRAGMA Actively Engages New Communities via PRAGMA Institutes and Mini-PRAGMA Workshops

• Water Disaster Management and Big Data
  — NII Shonan Meeting, 7 – 10 July 2014

• Mini-PRAGMA Indonesia, June 2013

• Lower Mekong Initiative Workshop
  — Hanoi August 18 – 22
Why 2X/year in-person working meetings

• Hosts highlight activities within their country
  – Chance to learn about research and activities that might otherwise go unnoticed
• Critical time for “shoulder-to-shoulder” collaborative work
• Recurring structure is a natural “clock” to keep distributed activity focused and moving forward
• Participation is open, but is based upon the desire to work with others.
• Virtual meetings, email, shared code repository are all practical methods for time between meetings
PRAGMA 26 – Theme Living with Big Data

Grad Students with Faculty Mentor

Interviewing Students during Lightning talks

Poster Session

Working Together
The Gap Between Technology Capability and Practice

• Technology developers are not always focused on the specific needs of domain scientists

• The pace of change can be dizzying

• Smaller groups of scientists are often left to fend for themselves

• PRAGMA focuses on bringing technologists and scientists together to make technology work in an international setting → Enable collaborative science
Enabling the Long Tail of **Team Science**

**Virtual Scientific Expeditions**

- Normal lake
- Lake during eutrophication
- Algae bloom

**Trust Envelope (network overlay)**

- PRAGMA Students
- PRAGMA Member Sites
- PRAGMA Community
Key Organizing Principle: Scientific Expeditions

Addressing Science Needs & Developing/Improving Tools

Savannah Burn 2006

NCHC SARS Task Force

Distribution of ash from March 2011: Tohoku Earthquake and Tsunami

- GEO Grid Disaster Task Force migrated services to other sites, including
  - NARL | NCHC | NSPO
  - UCSD | SDSC
  - OCC, Orkney, ERSDAC, NTT-data-CCS, CTC, Univ. Lille, and ITT
- Used aspects of virtualization porting and distributed resources at NCHC and SDSC

http://antisars.nchc.gov.tw/

Source: Fang-Pang Lin

http://disaster-e.geogrid.org
A “Resource” view of collaborative science

• Researchers want to work together
  – Identify essential data sets (what, where, how big? Allowable use models, ...)
    • Not all data is open
    • Not all data can be shared
  – Identify the types of resource(s) needed to accomplish analysis
    • Computing cluster, Visualization System, data handling, ...
    • If you are “lucky” these are all in one place
      – Most scientists are not so fortunate
Scientific Expeditions

- Information technology specialists + domain science application == Scientific Expeditions
- Domain scientists benefit from deep technical expertise
- Information technology specialists benefit from seeing how their tools/techniques can be used and improved.
Use “Overlay” Networks to provide a trusted environment for focused sharing of resources.
Some Specific PRAGMA Activities

Enabled International Collaboration

Applications
- LifeMapper Biodiversity
- Lake Ecology
- Network Experiment
- Cyber Learning

Technologies
- Data Provenance
- Overlay Networks
- Virtual Cluster Migration
- Application Integration

Students
Researchers
PRAGMA Helps Establish GLEON
From Scientific Expedition to Community Building on NCHC Ecogrid in Taiwan
8 Months: Concept to Deployment

- Science
- Need more than one lake to understand processes
P.C. Hanson and K.C. Weathers
GLEON Steering Committee co-Chairs

Mission: GLEON conducts innovative science by sharing and interpreting high-resolution sensor data to understand, predict and communicate the role and response of lakes in a changing global environment.
GLEON, a network of >100 lakes and ~$10^7$ data

- Technological challenges in sensing and information management
- Analytical challenges in interpreting and modeling very large data sets
- Scientific challenges in interpreting broad spatio-temporal gradients from many ecosystems
- Social and cultural challenges in diverse teams with diverse and distributed resources

A rich and productive research, education and outreach environment!
But most importantly a network of people

~ 450 individual members
~ 45 countries
Celebrating 12 Years

Metabolic changes and the resistance and resilience of a subtropical heterotrophic lake to typhoon disturbance

Jeng-Wei Tsai, Timothy K. Kratz, Paul C. Hanson, Nobuaki Kimura, Wen-Cheng Liu, Fang-Pan Lin, Hsiu-Mei Chou, Jiunn-Tzong Wu, and Chih-Yu Chiu

Lake-size dependency of wind shear and convection as controls on gas exchange

Jordan S. Read,1 David P. Hamilton,2 Ankur R. Desai,3 Kevin C. Rose,4,5 Sally MacIntyre,6 John D. Lenters,7 Robyn L. Smyth,5 Paul C. Hanson,8 Jonathan J. Cole,9 Peter A. Stach,10 James A. Rusak,11 Donald C. Pierson,12 Justin D. Brookes,13 Alo Laas,14 and Chin H. Wu1


Reconciling the temperature dependence of respiration across timescales and ecosystem types

Gabriel Yvon-Durocher, Jane M. Caffrey, Alessandro Cescatti, Matteo Dossena, Paul del Giorgio, Josep M. Gasol, José M. Montoya, Jukka Pumpanen, Peter A. Stach, Mark Trimmer, Guy Woodward & Andrew P. Allen
Nature 487, 472–476 (26 July 2012) doi:10.1038/nature11205

MACROSYSTEMS ECOLOGY
Creating and maintaining high-performing collaborative research teams: the importance of diversity and interpersonal skills

Kendra S Cheruvelil,2,7 Patricia A Soranno,2 Kathleen C Weathers3, Paul C Hanson1, Simon J Goring3, Christopher T Filstrup,9 and Emily K Read4

Large gradients of ecosystems enable new insights

Lakes help us understand fundamental principles of ecology.

How you form and manage collaborations matters!
PRAGMA-GLEON Expedition Goals

• Create a collaborative human and technological infrastructure that supports distributed team
• Apply a new hydrodynamic-water quality model, GLM-FABM-AED, to GLEON lakes.
• Consider a different set of rules that govern biological communities in lakes
• Expand research to additional lakes and broader research community

Leverage and expand science momentum...
Predicting Water Quality in Lakes
Developing Predictive Models using IPOP Overlay

- Lake eutrophication is a global issue, resulting in degraded water quality.
- Calibrate a new hydrodynamic model, check the model against data.
- Using IP-over-P2P (IPOP) to interconnect resources (and Virtual Machines) across multiple institutions, creating a "trust envelope".

Collecting light, temperature data

Cluster VPN in 3 Easy Steps
- Step 1: create a group on an XMPP server.
  - UF manages one — you can also deploy your own.

Paul Hanson, Craig Snortheim, Luke Winslow (U. Wisconsin), Cayelan Carey (Virginia Tech); Renato Figueiredo, Pierre St. Juste, Ken Subratie (U. Florida)
Biodiversity Expedition
Technology Trends Affecting Biodiversity

• Digitization of Collections
• Mobile Technologies
• Sensors – aquatic, terrestrial, and airborne
• Software Defined Networks

Lots of Data, Lots of Opportunities to Share

And Future Growth is Staggering

Today, the number of networked devices = the global population
By 2015, the number of networked devices = 2x the global population

2 M wingspan, 4.9 kg
Lifemapper – A Key Domain Science Tool

• Data library
  – Climate
    • Observed
    • IPCC Predicted Future Climate
  – Species
    • Occurrence Points
    • Potential habitat maps

• Tools
  – LmSDM: Species Distribution Modeling
  – LmRAD: Range and Diversity
LmSDM: Species Distribution Modeling

Species Occurrence Data

Environmental Data

Potential Habitat
LmRAD: Range and Diversity

Species Habitat Data

Presence Absence Matrix (PAM)

Multi-species analyses

Range and Diversity Quantifications
Biodiversity in Extreme Environments

Distribution Prediction by Sharing CI and Provenance Capture

Approach: Improve Lifemapper (LM)

• Extend previous LM work to enable data management and portability of software
• Increase the availability and flexibility of LM to enable scientists to
  • Assemble multi-species macro-ecology experiments
  • Perform other LM-facilitated data processing on:
    Unique datasets; Restricted use data
    Very large datasets

Aimee Stewart (Kansas), Nadya Williams, Philip Papadopoulos (UCSD), Reed Beaman (U Florida)
Antony van der Ent, Peter Erskinge (U Queensland)
Rimi Pepin (Sabah Parks)
UAV Flight (Kinabalu Donkey Ears)

Data Acquisition + Usual questions of who can access? Where can it be moved? How does it change over time? What related data products are available?
Expedition ➔ Infrastructure

• What are some of the key technologies and trends?

• How do we construct the systems needed without being restricted to a physical site?
Some Things that Happened on the Way to Cloud Computing

• Web Version 1.0 (1995)
• 1 Cluster on Top 500 (June 1998)
• Dot Com Bust (2000)
• Clusters > 50% of Top 500 (June 2004)
• Web Version 2.0 (2004)
• **Cloud Computing (EC2 Beta - 2006)**
• Clusters > 80% of Top 500 (Nov. 2008)
Gartner Emerging Tech - 2008

Big Data?
Gartner 2014

Figure 1. Hype Cycle for Emerging Technologies, 2014

- Cloud Computing
- Big Data
- Virtual Personal Assistants
- Articulating and Holographic Displays
- Software-Defined Anything
- Quantum Computing
- Human Augmentation
- Brain-Computer Interface
- Connected Home
- Virtual Reality
- NFC

Plateau will be reached in:
- less than 2 years
- 2 to 5 years
- 5 to 10 years
- more than 10 years
- obsolete before plateau

As of July 2014
What is fundamentally different about Cloud computing vs. Grid Computing

• Cloud computing – You *adapt the infrastructure* to your application
  – Should be less time consuming

• Grid computing – you *adapt your application* to the infrastructure
  – Generally is more time consuming

• Cloud computing has a financial model that seems to work – grid *never* had a financial model
System Virtualization – Underlying Technology for cloud computing

• Software the allows you to run a virtualized computer inside of a physical one

• Different systems create this software illusion
  – VMWare, Virtualbox, Xen, KVM,

• Different cloud systems give users a web-services interface to virtual systems
  – Amazon EC2, OpenStack, CloudStack, Eucalyptus, OpenNebula, ...
Physical Clusters $\rightarrow$ Virtual Clusters

• Beowulf and HPC Clusters are today the most common architectures for delivering scientific computing

• Specific Scientific Software can be quite complex to configure and maintain properly
  – This makes it difficult to move infrastructure from place to place
  – “Cloud computing” by itself doesn’t solve the underlying system issues.

• HPC Clouds are very close to reality
An HPC (cloud) for the 99%
(Production Date: 1 Jan 2015)

The next-generation of HPC will support high-performance virtual clusters @SDSC
(One reason why PRAGMA doesn’t focus running resources)
Celebrating 12 Years

Comet Will Serve the Long Tail of Science

CHALLENGES OUR PROPOSAL ADDRESSES
✓ Attract new users and communities
✓ Support diverse applications with complex workflows
✓ Ensure responsiveness for thousands of users
✓ Transfer, store, analyze, and share massive data sets
✓ Integrate with XSEDE

COMET COMPUTE SYSTEM

Cluster architecture
Fast standard nodes
Large-memory nodes
GPU-accelerated nodes
FDR InfiniBand

Storage architecture
Performance Storage
Durable Storage

Software
Science Gateways
Rich base of installed apps
Virtualization

USER & SYSTEM SUPPORT
New user orientation
XSEDE collaborations
FutureGrid

ALLOCATIONS & SCHEDULING
Optimized for throughput
Per-project allocation caps
Per-job core limits

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

XSEDE Service Providers
Internet2, ESnet @100G
(Universities, Labs)

Open Science Grid

Science Gateways

Cluster

Performance Storage

Durable Storage

UCSD Campus Bridging
(e.g., LHC Tier 2 Data Site)
Comet Architecture

Node-Local Storage
- 320 GB
- 72 HSWL
- N racks
- 72 HSWL
- N GPU
- 4 Large-Memory

Performance Storage
- 7 PB, 200 GB/s

Durable Storage
- 6 PB, 100 GB/s

FDR 36p
- 18 Mid-tier

IB Core
- (2x)

Bridge (4x)

Arista 40GbE
- (2x)

FDR 36p

FDR
- 72

FDR
- 72

FDR
- 72

FDR 36p
- 64
- 40GbE
- 128
- 10GbE

Internet 2
- R&E Network Access
- Data Movers

Additional Support Components
(not shown for clarity)
- NFS Servers, Virtual Image Repository,
  Gateway/Portal Hosting Nodes, Login
  Nodes, Ethernet Management Network,
  Rocks Management Nodes

> 45K Cores
High-Performance Virtualization on Comet

- Mellanox FDR InfiniBand HCAs with SR-IOV (Single Root IO Virtualization)
- Rocks and OpenStack Nova to manage VMs
- Flexibility to support complex science gateways and web-based workflow engines
  - Custom compute appliances and virtual clusters developed with FutureGrid and their existing expertise
  - Backed by virtualized Storage running over virtualized InfiniBand
Virtualization performance modest impact: Weather Modeling – 15% Overhead

- 96-core (6-node) calculation
- Nearest-neighbor communication
- Scalable algorithms
- SR-IOV incurs modest (15%) performance hit
- But still 20% faster*** than Amazon

WRF 3.4.1 – 3hr forecast

*** 20% faster despite SR-IOV cluster having 20% slower CPUs
Back to the Biodiversity Expedition ...  
... What do we need to do run Lifemapper in this type of environment?
Lifemapper Server Virtualization

Domain scientist’s viewpoint:
1. Extend previous Lifemapper work to enable data management (LmDBServer) and web services (LmWebServer) components virtualization
2. Increase the availability and flexibility of Lifemapper to enable scientists to
   - Assemble multi-species macro-ecology experiments
   - Perform other LM-facilitated data processing on:
     - Unique datasets
     - Restricted use data
     - Very large datasets

Cyber-infrastructure viewpoint:
1. Continue practical use of PRAGMA cloud infrastructure
2. Encapsulate the complexity of software build/configure in ROCKS rolls
3. Create a complete system as an end-to-end solution
4. Reduce cost of installing/configuring/replicating
Complexity of Scientific Applications
Lifemapper Server Roll – More than 20 Software components

Lifemapper
- LM code
  - Hdf44/hdf5
  - Subversion
  - Cmake
  - Byacc
  - Libraries
  - Gdal
  - Geos
  - Mod_python
  - SpatialIndex
  - Tiff
- LM data
  - Pgdg repo
  - Server
  - Client
  - Devel
  - Contrib
  - Openssl

Dependencies
- Postgresql
  - Pgdg repo
  - Server
  - Client
  - Devel
  - Contrib
  - Openssl
- Postgis
  - Geos
  - Proj
  - Pgbouncer
- Python modules
  - Cheetah
  - Cherrypy
  - Cython
  - Psycopg2
  - Pylucene
  - Rtree
- Mapserver
  - Elgis repo
  - Vera fonts
  - Fribidi

Configuration
- Postgresql
- Postgis
- Pgbouncer
- LM components

Total: 56 RPMs (packages)
What we are trying to do

• Increase availability and flexibility of Lifemapper Server as a complete system
  – reduce cost of installing/configuring/replicating and ease burden of integrating hardware and software
• Enable a fast “workflow” from software update to server availability:
  – Minimize time spent on software build and configuration
  – Automate most hands-on tasks.
  – Essential: have test cases for all installed components and their configuration
• Prepare for greater quantity and quality of data and complexity of operations
  – From low resolution climate data to high resolution satellite imagery for Mt. Kinabalu
  – From simple single-species SDM experiments to multi-species macro-ecology experiments with more species

This work is a part of PRAGMA’s “Resources and Data” working group

Lifemapper
- https://lifemapper.org
- https://github.com/lifemapper/
Rocks
- http://www.rocksclusters.org
Pragmagrid GitHub
- https://github.com/pragmagrid/lifemapper
- https://github.com/pragmagrid/lifemapper-server
PRAGMA-Developed Tech

• You have a Virtual Cluster with your scientific software properly configured, What’s the next step?

• PRAGMA_boot
  – Practically solve the problem of moving a virtualized cluster from one cloud hosting system to another
Virtual Cluster Image

• Define a standard way to share cluster images
  – E.g. frontend: LmDbServer, compute: LmCompute
Deployment

• Different hosting environments:
  – UCSD uses Rocks Clusters
  – AIST (Japan) uses OpenNebula
  – ...

• How can deploy the Virtual Cluster Image?

pragma_boot:
https://github.com/pragmagrid/pragma_boot
The *pragma_boot* script

*pragma_boot* is the main program to instantiate Virtual Machine in Pragma. It accepts the following arguments:

- **--list** list the available images
- **--num_cpus N** the number of compute node to start up (default to 0)
- **--vcname vcname** the name of the virtual cluster to start up (the name must be in the database)
- **--base_path path** the base path of the VM database
- **--key path** The ssh key that will be authorized on the frontned of the cluster (default is /root/.ssh/id_rsa.pub)

*pragma_boot* invokes the following subscripts which will be invoked in the order described below. In the commands below the *ve_driver* will be replaced with the local Virtual Engine (VE) driver (the base path used to find all the VE drivers can be configured in the file site_conf.conf) *site_conf.conf* should be used also to set the path for the temporary_directory used for staging all VM images.

- **ve_driver/fix_images** prepare the given VC image to be run on the current system (fix kernel, drivers, boot options, for current platform, etc.). It's input arguments are (in the following order):
  i. **vc_in_file** the path to the vc-in.xml file of the virtual machine we have to convert
  ii. **temp_directory** the temporary directory used to place all the temporary virtual
  iii. **node_type** a command separated list of node type to be prepared (e.g. "frontend,compute")

- **ve_driver/allocate** this script takes care of verifying that there are enough resources to satisfy the user request, if so it will also allocate public IP, private IPs, MAC addresses, and computing resources. If the system can create SMP nodes it can allocate less compute node with multiple cpus in each node. If successful it will write a /root/vc-out.xml file inside the various virtual machines images (see below for more info)
  i. **num_cpus** it specifies the number of CPU requested by the user.
  ii. **vc_in_path** it points to the vc-in.xml of the selected cluster
  iii. **vc_out_path** this should point to the path where the frontend vc-out.xml will be saved
Rocks Implementation

1. Format conversion (e.g. Xen->KVM or raw qCOW)

2. Assign local resources for guest cluster (network, hosts, disks, etc)

3. Turn on Virtual Cluster

Virtual Cluster Image
- Frontend Image.gz
- Compute Image.gz

vc-in.xml

fix_image

allocate

vc-out.xml

boot

Comput images
- Compute Image
- Frontend Image

Assign local resources for guest cluster (network, hosts, disks, etc)
Next logical Step

1. Scientific Software Installed in Virtual Cluster (VC)
2. Can boot virtual cluster on a variety of hosting resources (without redefining your VC)
3. How do you get to protected/sensitive/unpublished data?
We Can move VCs between different clouds, can we get controlled access to remote data?
VC + Pragma_boot + Overlay network + Data
Source: This is a lot to put together

Is there something we can do to make it simpler for the end-user?
Personal Cloud Controller (PCC)
(Yuan Luo (IU), Shava Smallen (UCSD), Beth Plale (IU), Philip Papadopoulos(UCSD))

- Goals:
  - Enable **lab/group** to easily **manage** application **virtual clusters** on available resources
  - Leverage PRAGMA Cloud tools: `pragma_bootstrap`, IPOP, ViNE.
  - Lightweight, extends HTCondor from U Wisc.
  - Provide command-line and Web interfaces

- Working Group: Resources
Personal Cloud Controller (PCC) - cont.

• Current status
  – Start and monitor virtual cluster using pragma_bootstrap via HTCondor (VM GAHP)
  – Web interface prototype (PHP)

• Near-term goals
  – Add increased controllability and robustness (April – June)
  – Multi-site clusters (July – Sept)

• Longer-term goals
  – Data-aware scheduling
  – Fault tolerance
  – Provenance
A Prototype Web Interface

Figure 2: Frontend web interface of PCC showing the launch of a virtual cluster.
Shared Software Development

- Experimental Network Testbed
- Virtual Cluster Migration
- Network Overlays
- LifeMapper in a virtual machine
PRAGMA Works Because of a Culture of Sharing and Trust

- Mutual technology interests – Some shared development.
- Mutual scientific interests
- Bridging the gap between technology capabilities and domain science needs
- Open to experiments in science and technology
- Next Meeting: Bloomington, Indiana Oct 15-17