Preparing Scientists for Scalable Software Development

Valerie Maxville
Education Program Leader

About iVEC
• Five members:
  – Four public universities
  – CSIRO
• Three facilities
• Education and Industry Programs
• State and federal government funding

iVEC Provides eResearch Infrastructure
a.k.a. eScience infrastructure or Cyber-infrastructure

• Discipline and project affect eResearch needs
• Components of each project may include:
  – Shared large, expensive instruments
  – High performance computing
  – Scientific visualisation
  – High speed networks
  – Collaboration tools, e.g. Access Grid
  – Shared data repositories
  – Web portals (access resources and share information)
• Researchers choose a palette of components

Federal Funding for eResearch
• Through NCRIS, the Government is providing $542 million over 2005-2011 to provide researchers with major research facilities, supporting infrastructure and networks necessary for world-class research.
  – Evolving Bio-Molecular Platforms and Informatics
  – Integrated Biological Systems
  – Characterisation
  – Fabrication
  – Biotechnology Products
  – Networked Bioscience Framework
  – Optical and Radio Astronomy
  – Integrated Marine Observing System
  – Platforms for Collaboration
  – Terrestrial Ecosystems Research Network
  – Population health and clinical data linkage
**Major Supported Projects**

- Computational Nanoscience (NRI)
- Computational Geology (Minerals Downunder, ACcESS, SEEGrid, AuScope)
- Radioastronomy (eVLBI, ASKAP)
- Computational Biology and Biochemistry
- Bioinformatics (Murdoch and UWA)
- Water Research (UWA)
- Theoretical Physics (Curtin)

... and many more ...

**Scale of SKA**

- An international consortium representing more than 15 countries including Europe, USA, Australia, Canada, China, India and South Africa
- €300 M for Phase 1
- €1,200 M for Phase 2
- The third phase is less well-defined
- Operational in 10-15 years, 70 year lifespan

**New $80 million SKA Science Centre**

Federal Budget announcement 12th May 2009

“Australian National Centre of SKA Science to be established in Perth to host new high performance computing facilities and expertise to support SKA research and other high-end science”

The Centre will be part of iVEC

**What does this mean for researchers?**

- The research lifecycle is changing – an evolution rather than a sea-change
- Bigger and more complex problems require new methodologies and relationships
- Policy and funding are increasingly dictating collaboration
- Advanced networks are essential
- It’s more about data than technology
- Many social and organisational factors

*IVEC is responsible for providing resources and developing capacity for Western Australia*
Preparing Scientists for Scalable Software Development

SECSE 09 Workshop

Scale issues for iVEC

- Increased number of users
- Increase in use per user
- More diversity in disciplines
- More diversity in user background
- More resources on offer, in some combination:
  - Multiple compute resources
  - Scientific visualisation
  - Large scale data storage
  - Collaborative tools and workspaces
  - High bandwidth networking

Scale issues for Developers

- Number of Developers
- Number of Users
- Colocated or distributed teams
- Duration of project
- KLOC
- Multicore / parallel code
- Multiple target platforms
- Outward Dependencies
- Inward Dependencies

Characterising eResearch Projects

<table>
<thead>
<tr>
<th>Scale Indicators</th>
<th>CC1</th>
<th>CC2</th>
<th>CC3</th>
<th>CC4</th>
<th>CC5</th>
</tr>
</thead>
<tbody>
<tr>
<td># Developers</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Users</td>
<td>Ext</td>
<td>Ext</td>
<td>Ext</td>
<td>Ext</td>
<td>Ext</td>
</tr>
<tr>
<td>Colocated</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Duration</td>
<td>10+</td>
<td>10+</td>
<td>5+</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>KLOC</td>
<td>100</td>
<td>320</td>
<td>200</td>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td>Multicore</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Multiplatform</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Outward Dependencies</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Inward Dependencies</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

SWEBOK and Selected Chemistry Projects

<table>
<thead>
<tr>
<th>Knowledge Area</th>
<th>CC1</th>
<th>CC2</th>
<th>CC3</th>
<th>CC4</th>
<th>CC5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requirements</td>
<td>A-V</td>
<td>A-V</td>
<td>A-V</td>
<td>A-V</td>
<td>A-V</td>
</tr>
<tr>
<td>Design</td>
<td>N-V</td>
<td>A-V</td>
<td>A-V</td>
<td>A-V</td>
<td>A-V</td>
</tr>
<tr>
<td>Construction</td>
<td>N-V</td>
<td>A-V</td>
<td>A-V</td>
<td>N-I</td>
<td>N-I</td>
</tr>
<tr>
<td>Testing</td>
<td>N-V</td>
<td>A-V</td>
<td>A-V</td>
<td>N-I</td>
<td>N-I</td>
</tr>
<tr>
<td>Maintenance</td>
<td>N-V</td>
<td>A-V</td>
<td>A-V</td>
<td>A-V</td>
<td>A-V</td>
</tr>
<tr>
<td>Configuration Mgmt</td>
<td>A-V</td>
<td>A-V</td>
<td>A-V</td>
<td>A-V</td>
<td>A-V</td>
</tr>
<tr>
<td>Engineering Mgmt</td>
<td>N-V</td>
<td>A-V</td>
<td>A-V</td>
<td>N-I</td>
<td>N-I</td>
</tr>
<tr>
<td>Engineering Process</td>
<td>N-V</td>
<td>N-V</td>
<td>N-V</td>
<td>N-V</td>
<td>N-V</td>
</tr>
<tr>
<td>Tools and Methods</td>
<td>N-V</td>
<td>N-V</td>
<td>N-V</td>
<td>N-V</td>
<td>N-V</td>
</tr>
<tr>
<td>Quality</td>
<td>N-V</td>
<td>A-V</td>
<td>A-V</td>
<td>N-I</td>
<td>N-I</td>
</tr>
</tbody>
</table>

(A = Applied, N = Not applied) - (V = Valuable, I = Inappropriate)
Preparing Scientists for Scalable Software Development

SECSE 09 Workshop

AuScope

Integration of Observations, Mathematics and Numerical Simulation

© CSIRO 2003

Data Collection: Virtual Core Library

A variety of interfaces to suit user capabilities

Education Program

- Identify and drive development of curricula for advanced computing across the IVEC partnership
- Publicise and oversee the IVEC internship program
- Encourage e-Research and the use of advanced computing in the IVEC partners
- Organise networking events for eResearch community

AuScope Users

Compute Grid

Data Grid

APAC

AuScope Grid

Hardware Software

Elwood TB Maintenance

Visualisation Tools

3D Modelling Tools

Petroleum Industry

SAPAC

TPAC

VPAC

AC3

IVEC

Environment Industry

Minerals Industry

Insurance

Expert Scientist

Non-Expert Scientist

University Student

General Public

School Students
Preparing Scientists for Scalable Software Development

Building eResearch capacity

iVEC Wish List

• If you could choose the things you would most like our users to know… what would they be?

• Mix in "basic training" in eResearch… and a dash of Software Engineering…

Education Program 2007

• Free, full-day courses:
  • Introduction to iVEC
  • Programming with MPI
  • MPI Applications and Optimisation

• Intern program:
  • 12 weeks, $8000 tax free
  • 1 week induction - parallel programming
  • Loose guidance
  • Final presentation and report

Intern Induction

Education Program 2009

• Tailored 1 hour presentation:
  • Introduction to IVEC

• Free courses:
  • Introduction to Linux
  • Developing Parallel Applications
  • Scaling Parallel Applications
  • Scaling Software Development

• Courses under development:
  • Data management
  • Grid computing and web services
  • Visualisation

Education Program 2009

• Intern program:
  • 12 weeks - $6000 tax free
  • 1 week induction - wide range of topics
  • Collaborative workspace - wiki, Sakai
  • Weekly reports
  • Regular group meetings
  • 1/2 day presentation workshop - paper, podcast, presentation
  • Final presentation at IVEC eResearch Forum
Preparing Scientists for Scalable Software Development

**Depth of Knowledge / Target User Base**

<table>
<thead>
<tr>
<th>SWEBOK Knowledge Area</th>
<th>Individual</th>
<th>Group</th>
<th>Public</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requirements</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>Design</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>Construction</td>
<td>+</td>
<td>++</td>
<td>+++</td>
</tr>
<tr>
<td>Testing</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>Maintenance</td>
<td>+</td>
<td>++</td>
<td>+++</td>
</tr>
<tr>
<td>Configuration Mgmt</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>Engineering Mgmt</td>
<td>+</td>
<td>++</td>
<td>+++</td>
</tr>
<tr>
<td>Engineering Process</td>
<td>+</td>
<td>++</td>
<td>+++</td>
</tr>
<tr>
<td>Tools and Methods</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>Quality</td>
<td>+</td>
<td>++</td>
<td>+++</td>
</tr>
</tbody>
</table>

* **light coverage** ++ **deeper coverage** +++ **formal**

**Certificate in Advanced Computing**

- **Advanced Computing I**
  - concepts in advanced computing, choosing solutions, parallel programming, requirements, design, testing, configuration management, tools and methods
  - Information Management
  - storing research data, metadata, standards, compliance, tools
- **Advanced Computing II**
  - algorithm analysis, concurrency, design, testing, tools

**Early Feedback - Diversity**

- Even distribution - non-user, new user, longer term user (> 6 months)
- Mre self-taught programmers than those with computing qualifications
- Diverse domains, needs, project size, project maturity

**Early Feedback - Delivery**

- Delivery preferences:
  1. Online: interactive
  2. At institution OR Online: presentation
  3. Central location OR Online: podcast
- Full day courses preferred

**Feedback - Content**

- **Additional:**
  - All want visualisation and grid
  - Most want data management and web
  - Few want project management
- **Languages:**
  - Most: Python, MPI
  - Many: OpenMP
  - Some: Linux, FORTRAN

- **Most Useful:**
  - Optimising MPI
  - Using iVEC (compiling and queues)
  - Concepts in Parallel
  - Design and algorithms
  - Reuse and libraries
  - Project management
Open Issues in Education

- Balance between general/niche content
- Avoid competing with computing faculties
- Should it be almost “In Case of Emergency” => do this course? (on demand)
- Do they know what they need?
- From ICSE and associated workshops/tutes:
  - Patterns - Testing
  - Tools - Social/Cultural issues
- Also:
  - Architecture - Solution choice
  - Mash-ups - Complexity and memory
  - Astroinformatics - Robustness (checkpointing)
  - Project lifecycles (similar to OSS?)

Acknowledgements

- Course participants
- iVEC: Dr Daniel Grimwood
  Dr Chris Hines
- AuScope: Ryan Fraser
  Terry Rankine