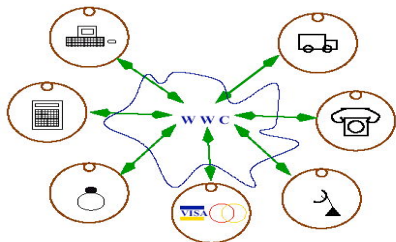


Enabling Synchronous Computation on Volunteer Computing Grids

Carlos Varela, cvarela@cs.rpi.edu
Rensselaer Polytechnic Institute
<http://wcl.cs.rpi.edu/>

Faculty	Graduate Students
Malik Magdon-Ismail, CS	Nathan Cole, Travis Desell
Heidi Newberg, AstroPhys	Kaoutar El Maghraoui, Ph.D., 2007
Bolek Szymanski, CS	Wei-Jen Wang, Ph.D., 2006



INRIA BOINC Workshop
Grenoble, France
September 11, 2008

Research Challenges/Directions

- **Applications:**

Astroinformatics, particle physics, bioinformatics, virtual surgical planning, fluid dynamics, aeronautical design, climate modeling, etc.

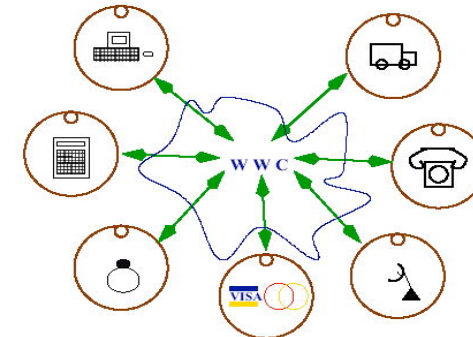
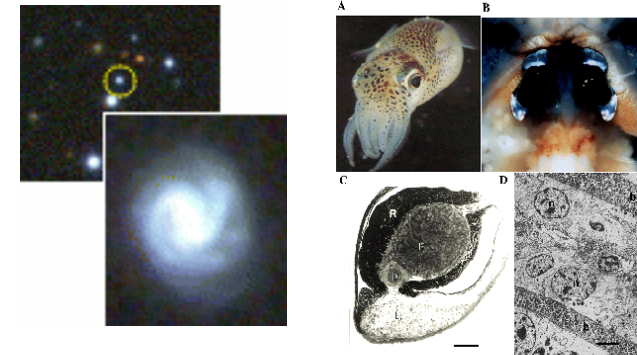
- **Middleware bridging the gap:**

Decentralized coordination; scheduling with inaccurate, partial knowledge about distributed resources; fault-tolerance; security; programmability; high performance

- ✓ **Programming/coordination abstractions/models/languages.**
- ✓ **Decentralized resource management protocols.**
- ✓ **Integrated static and dynamic analyses techniques for performance prediction and evaluation.**
- ✓ **Heterogeneity-tolerant failure-oblivious algorithms.**

- **Infrastructure:**

Supercomputer centers, virtual grids, campus grids, grids in a box, planetary-scale networks



Milky Way Origin and Structure

- **Problem Statement:**

What is the structure and origin of the Milky Way galaxy?

How to analyze data from 10,000 square degrees of the north galactic cap collected in five optical filters over five years by the Sloan Digital Sky Survey? (over 10Tb in images)

- **Applications/Implications:**

Astrophysics: origins and evolution of our galaxy.

- **Approach:**

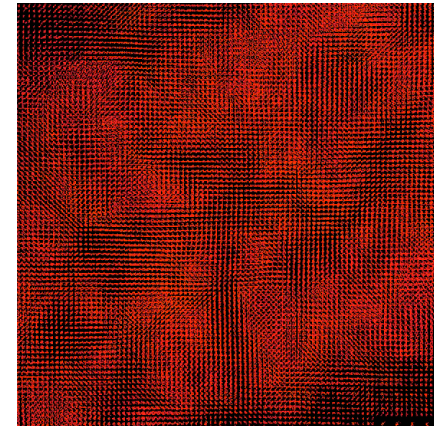
Experimental data analysis and simulation

To use photometric and spectroscopic data for millions of stars to separate and describe components of the Milky Way

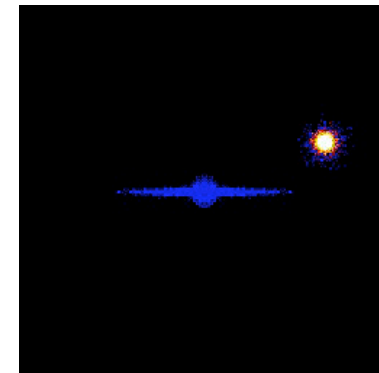
- **Software:**

Generic Maximum Likelihood Evaluation (GMLE) framework.

MilkyWay@Home BOINC project.



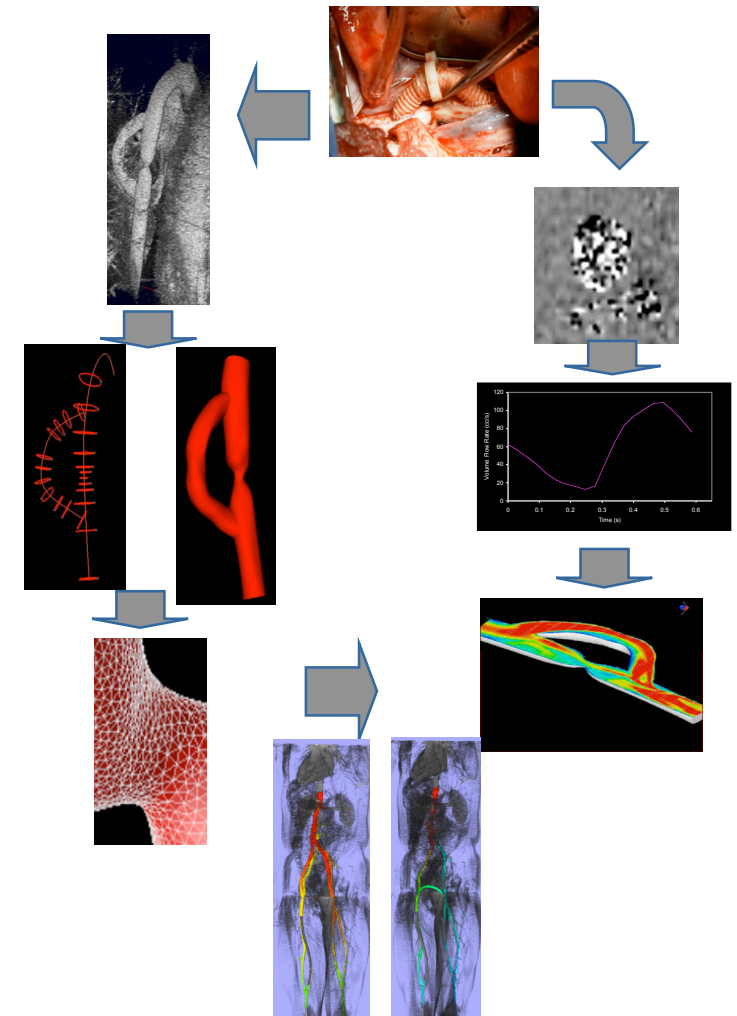
Ben Moore, Inst. Of
Theo. Phys., Zurich



Kathryn V. Johnston, Wesleyan
Univ.

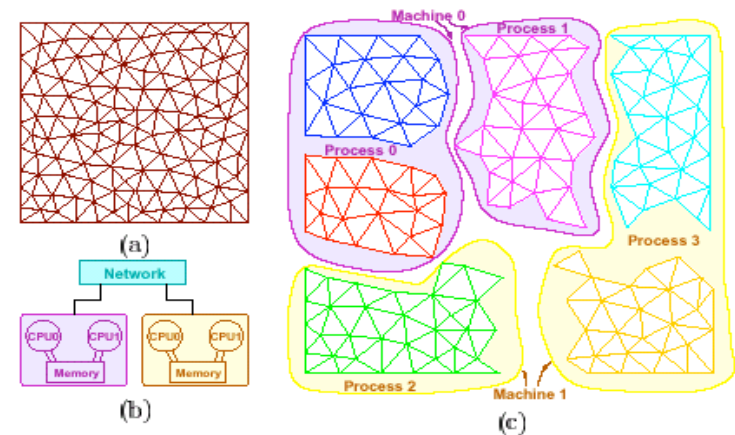
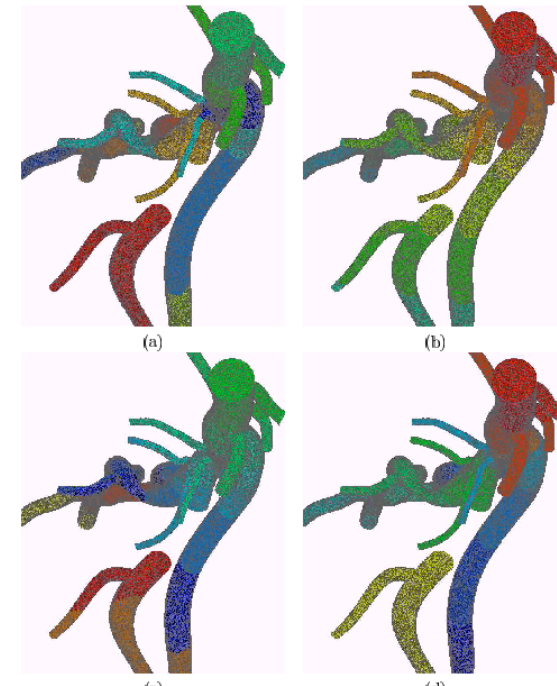
Virtual Surgical Planning

- **Investigators:**
K. Jansen, M. Shephard (RPI),
C. Taylor, C. Zarins (Stanford)
- **Problem Statement:**
How to develop a software framework to enable virtual surgical planning based on real patient data?
- **Applications/Implications:**
Surgeons will be able to virtually evaluate vascular surgical options based on simulation rather than intuition alone.
- **Approach:**
Scan of real patient is processed to extract solid model and inlet flow waveform.
Model is discretized and flow equations solved.
Multiple alterations to model are made within intuitive human-computer interface and evaluated similarly.
- **Software:**
MEGA (SCOREC discretization toolkit)
PHASTA (RPI flow solver).



Adaptive Partial Differential Equation Solvers

- **Investigators:**
J. Flaherty, M. Shephard B. Szymanski, C. Varela (RPI)
J. Teresco (Williams), E. Deelman (ISI-UCI)
- **Problem Statement:**
How to dynamically adapt solutions to PDEs to account for underlying computing infrastructure?
- **Applications/Implications:**
Materials fabrication, biomechanics, fluid dynamics, aeronautical design, ecology.
- **Approach:**
Partition problem and dynamically map into computing infrastructure and balance load.
Low communication overhead over low-latency connections.
- **Software:**
Rensselaer Partition Model (RPM)
Algorithm Oriented Mesh Database (AOMD).
Dynamic Resource Utilization Model (DRUM)



RPI Computational Center for Nanotechnology Innovations (CCNI)

Rensselaer



CCNI Computational Center for Nanotechnology Innovations

RENSELAER POLYTECHNIC INSTITUTE



[News Releases](#)

[CCNI Fact Sheet](#)

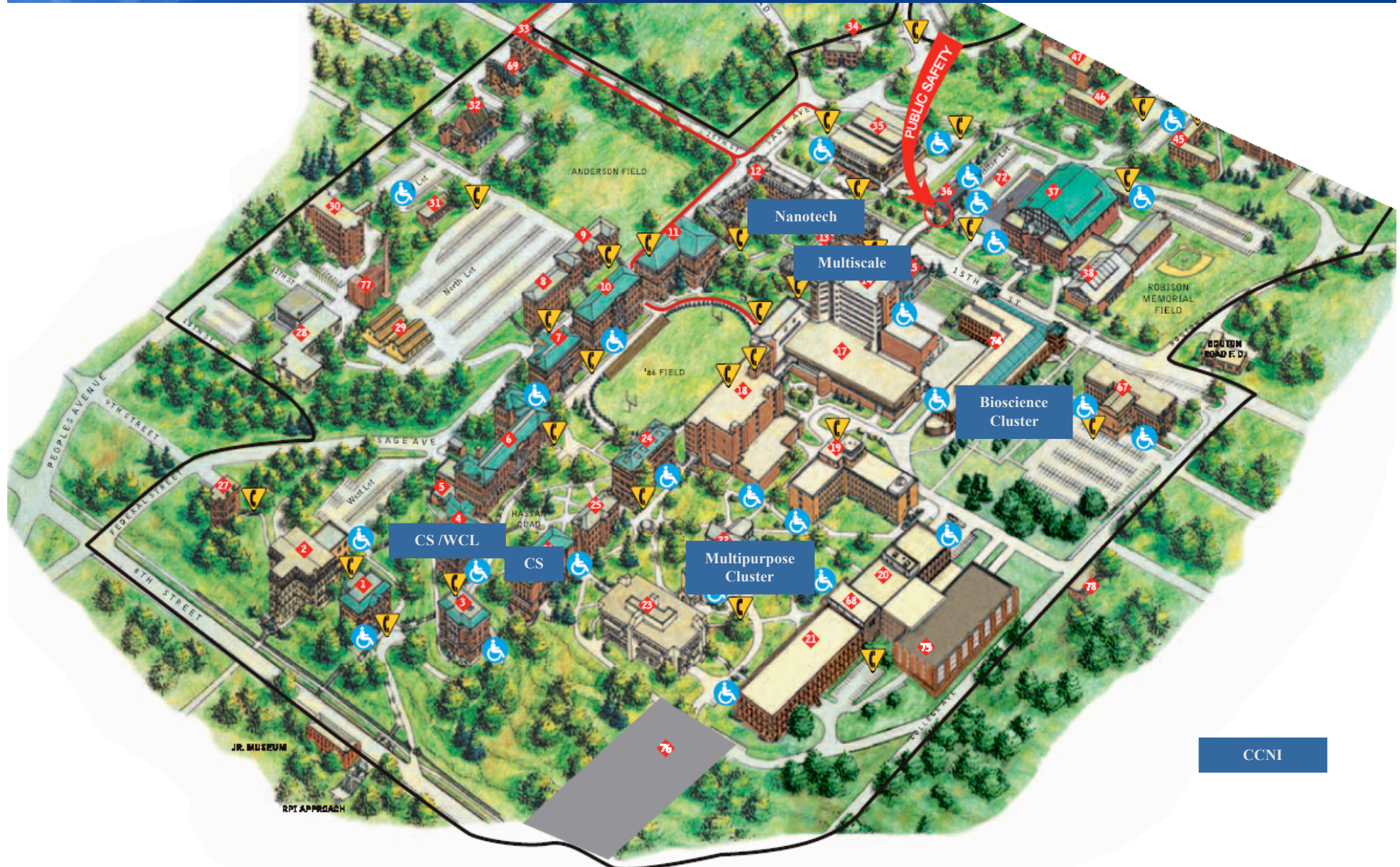
[Economic Impact](#)

[CCNI Facilities](#)

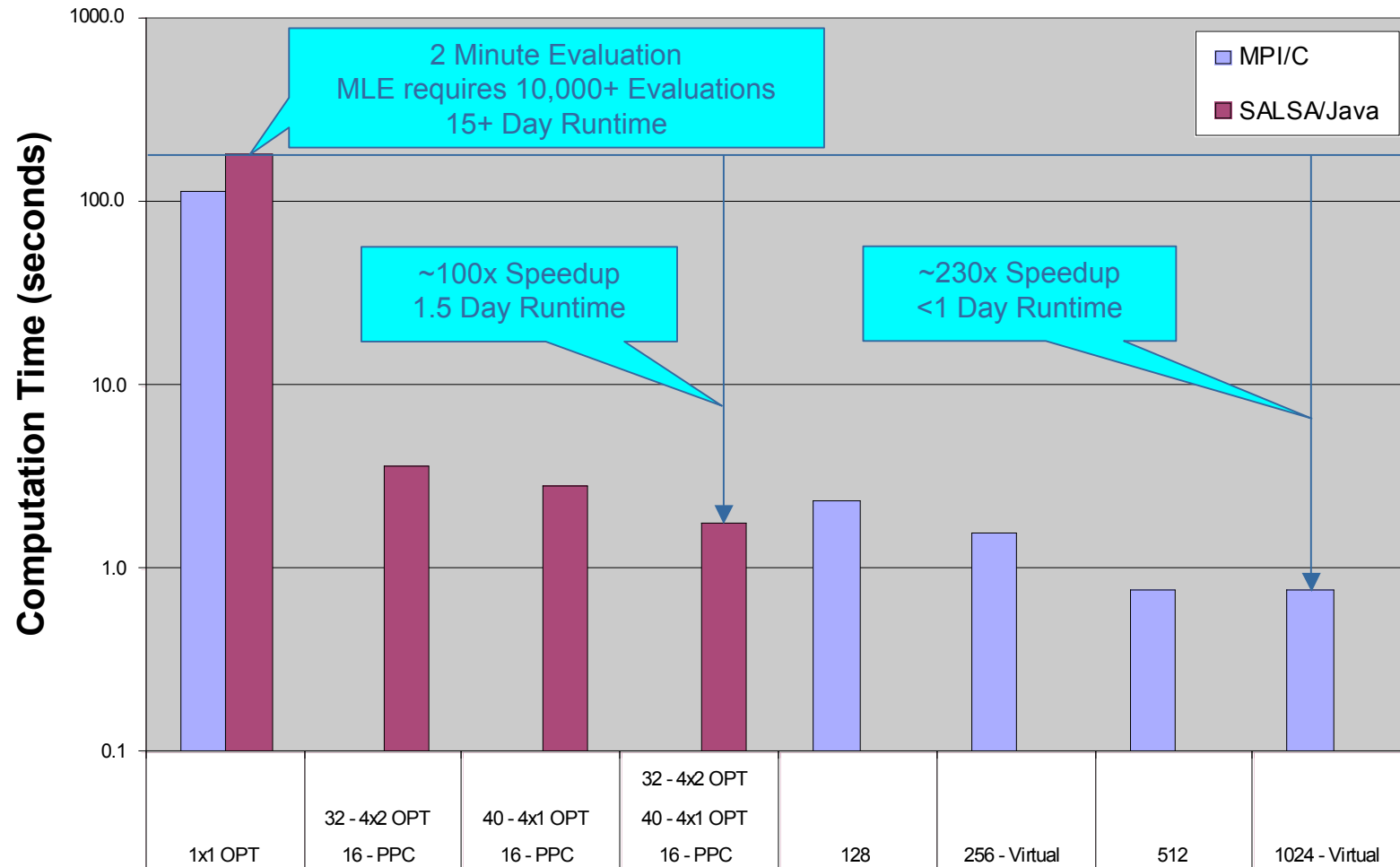
June 2007
Top500 list

The World's Most Powerful University-Based Supercomputing Center

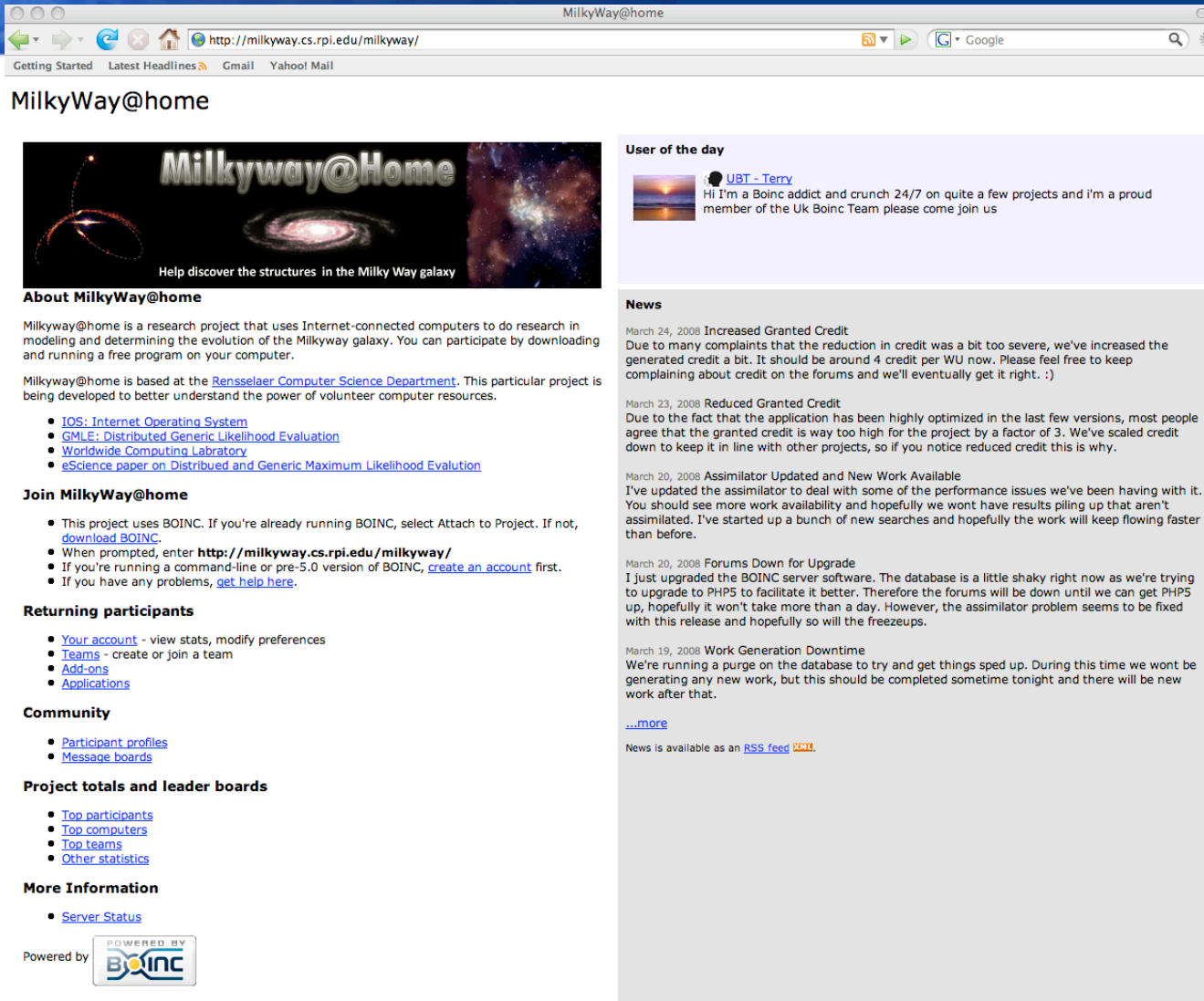
Map of Rensselaer Grid Clusters



Maximum Likelihood Estimation on RPI Grid and BlueGene/L Supercomputer



MilkyWay@Home: Volunteer Computing Grid



The screenshot shows a web browser window with the address bar containing <http://milkyway.cs.rpi.edu/milkyway/>. The page title is "MilkyWay@home". The browser's address bar also shows "Google" and a search icon. The page content includes a navigation menu with "Getting Started", "Latest Headlines", "Gmail", and "Yahoo! Mail". The main content area features a banner image with the text "Milkyway@Home" and "Help discover the structures in the Milky Way galaxy". Below the banner is an "About MilkyWay@home" section, followed by a "User of the day" section featuring a profile for "UBT - Terry". The "News" section contains several entries dated March 2008, including "Increased Granted Credit", "Reduced Granted Credit", "Assimilator Updated and New Work Available", "Forums Down for Upgrade", and "Work Generation Downtime". The page also includes sections for "Join MilkyWay@home", "Returning participants", "Community", "Project totals and leader boards", and "More Information". At the bottom left, there is a "Powered by BOINC" logo.

MilkyWay@home

Help discover the structures in the Milky Way galaxy

About MilkyWay@home

Milkyway@home is a research project that uses Internet-connected computers to do research in modeling and determining the evolution of the Milkyway galaxy. You can participate by downloading and running a free program on your computer.

Milkyway@home is based at the [Rensselaer Computer Science Department](#). This particular project is being developed to better understand the power of volunteer computer resources.

- [IOS: Internet Operating System](#)
- [GMLE: Distributed Generic Likelihood Evaluation](#)
- [Worldwide Computing Laboratory](#)
- [eScience paper on Distributed and Generic Maximum Likelihood Evaluation](#)

Join MilkyWay@home

- This project uses BOINC. If you're already running BOINC, select Attach to Project. If not, [download BOINC](#).
- When prompted, enter <http://milkyway.cs.rpi.edu/milkyway/>
- If you're running a command-line or pre-5.0 version of BOINC, [create an account](#) first.
- If you have any problems, [get help here](#).

Returning participants

- [Your account](#) - view stats, modify preferences
- [Teams](#) - create or join a team
- [Add-ons](#)
- [Applications](#)

Community


- [Participant profiles](#)
- [Message boards](#)

Project totals and leader boards


- [Top participants](#)
- [Top computers](#)
- [Top teams](#)
- [Other statistics](#)

More Information

- [Server Status](#)

Powered by 

User of the day

 [UBT - Terry](#)
Hi I'm a Boinc addict and crunch 24/7 on quite a few projects and i'm a proud member of the Uk Boinc Team please come join us

News

March 24, 2008 **Increased Granted Credit**
Due to many complaints that the reduction in credit was a bit too severe, we've increased the generated credit a bit. It should be around 4 credit per WU now. Please feel free to keep complaining about credit on the forums and we'll eventually get it right. :)


March 23, 2008 **Reduced Granted Credit**
Due to the fact that the application has been highly optimized in the last few versions, most people agree that the granted credit is way too high for the project by a factor of 3. We've scaled credit down to keep it in line with other projects, so if you notice reduced credit this is why.

March 20, 2008 **Assimilator Updated and New Work Available**
I've updated the assimilator to deal with some of the performance issues we've been having with it. You should see more work availability and hopefully we won't have results piling up that aren't assimilated. I've started up a bunch of new searches and hopefully the work will keep flowing faster than before.

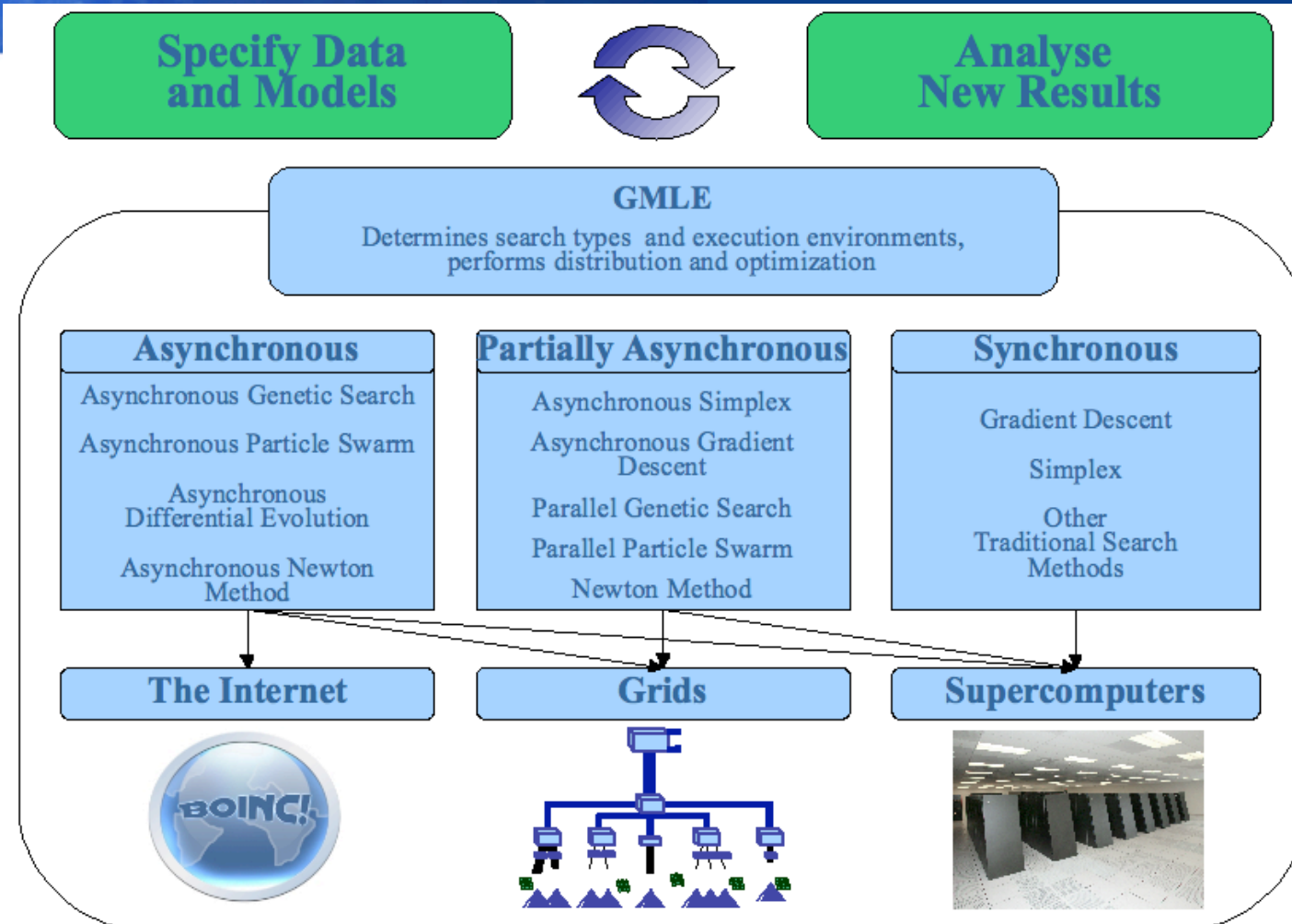
March 20, 2008 **Forums Down for Upgrade**
I just upgraded the BOINC server software. The database is a little shaky right now as we're trying to upgrade to PHP5 to facilitate it better. Therefore the forums will be down until we can get PHP5 up, hopefully it won't take more than a day. However, the assimilator problem seems to be fixed with this release and hopefully so will the freezeups.

March 19, 2008 **Work Generation Downtime**
We're running a purge on the database to try and get things sped up. During this time we won't be generating any new work, but this should be completed sometime tonight and there will be new work after that.

[...more](#)

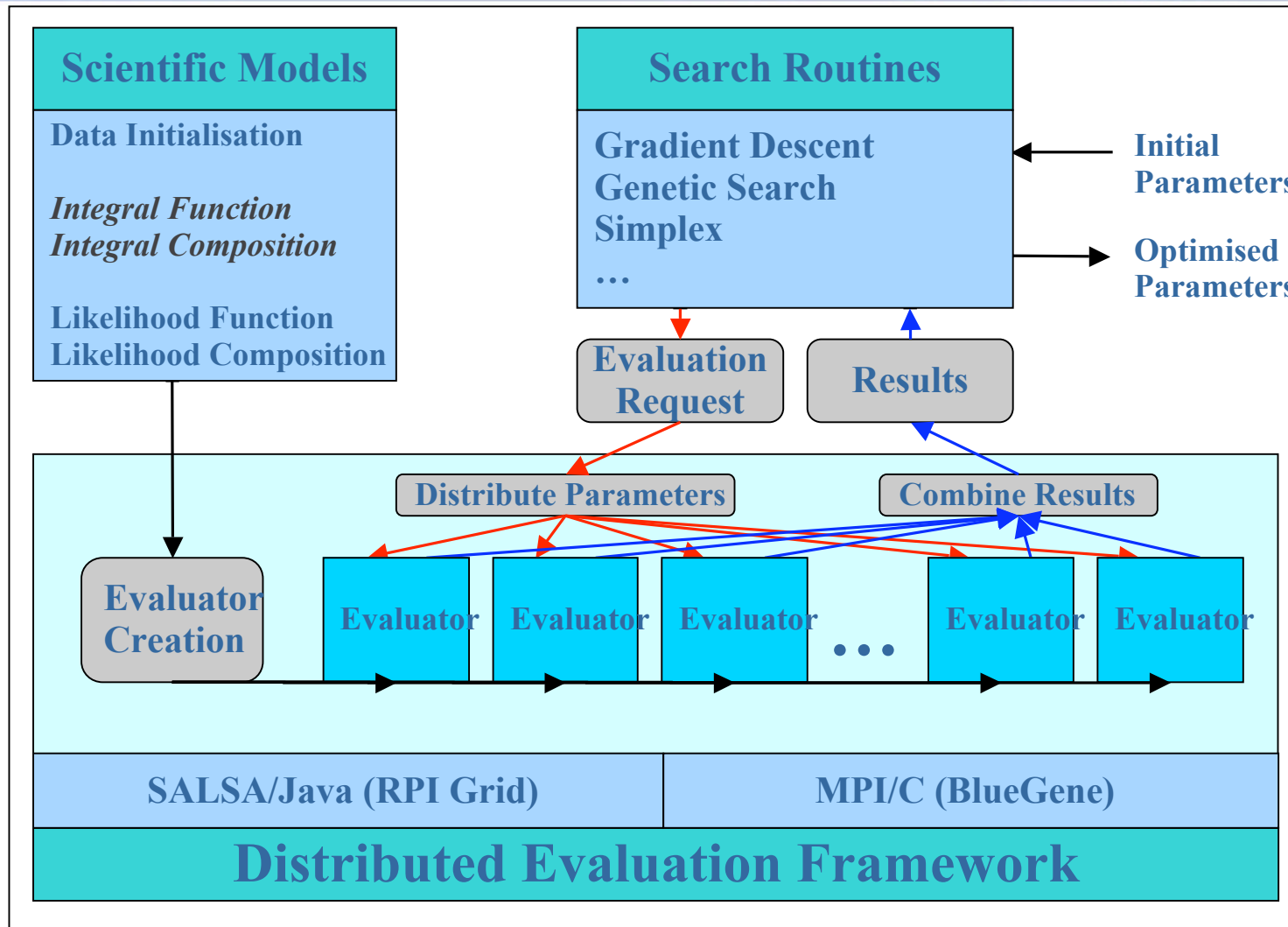
News is available as an [RSS feed](#) 

Generic Maximum Likelihood Estimation



T. Desell, N. Cole, M. Magdon-Ismail, H. Newberg, B. Szymanski and C. Varela, "Distributed and Generic Maximum Likelihood Evaluation", *eScience 2007*, Best Paper (Finalist) Award.

Synchronous Software Architecture



Asynchronous Search Strategies

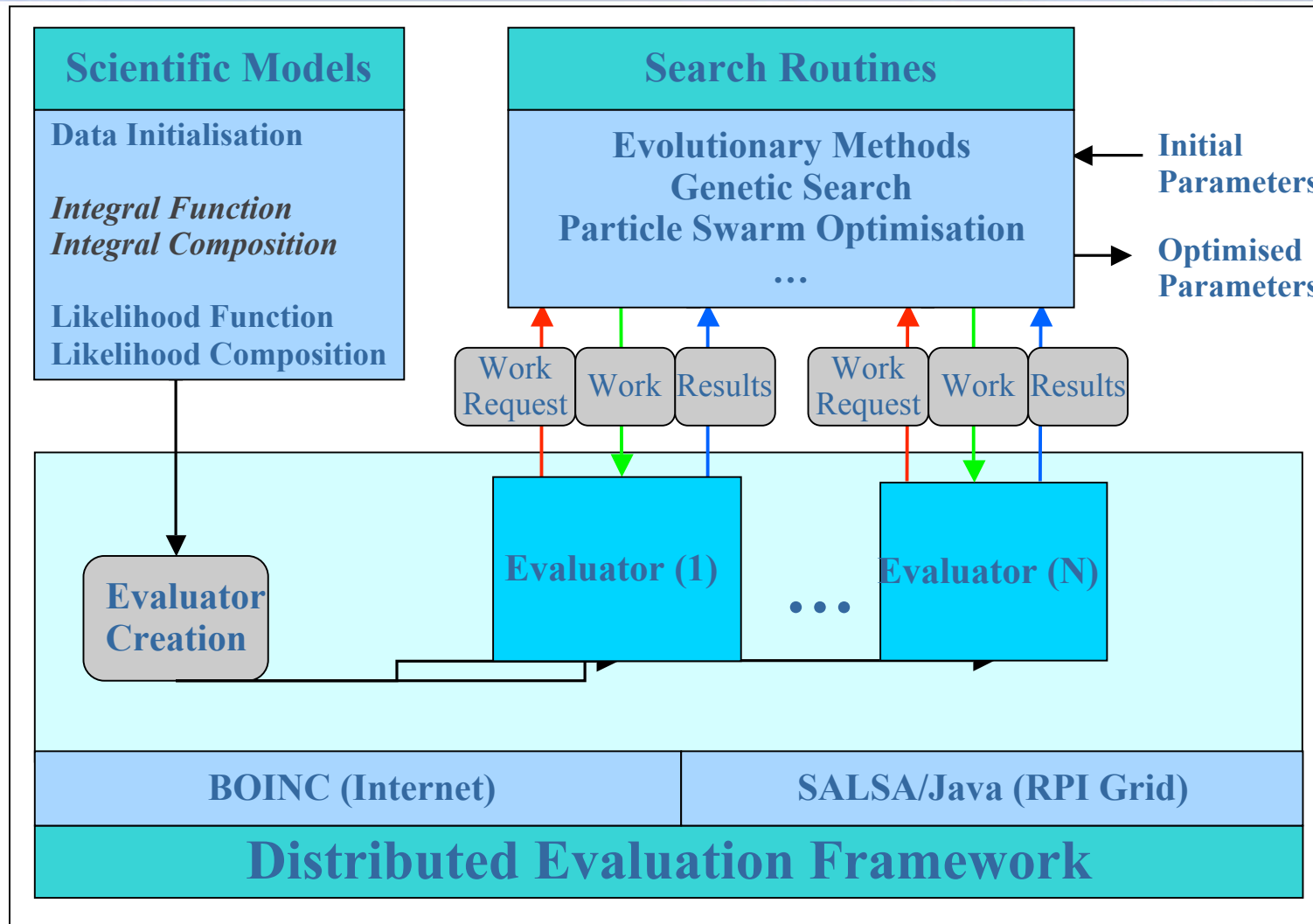
- Use an asynchronous search methodology
 - No explicit dependencies
 - No iterations
- Continuously updated population
 - N individuals are generated randomly for the initial population
 - When a evaluator requests more work, stochastically apply operators to members of the population to create new parameter sets to evaluate
 - When results arrive, update the population

T. Desell, B. Szymanski and C. Varela,

“Asynchronous Genetic Search for Scientific Modeling on Large-Scale Heterogeneous Environments”, *Heterogeneous Computing Workshop (IPDPS) 2008*.

“Asynchronous Hybrid Genetic-Simplex Search for Modeling the Milky Way Galaxy using Volunteer Computing”, In *Genetic and Evolutionary Computation Conference (GECCO) 2008*.

Asynchronous Software Architecture



Synchronous Volunteer Computing Challenges

- **Security**
 - Clients need to communicate with each other
 - A malicious client gaining control over other clients could cause a DDOS attack or damage other clients
- **Correctness**
 - One bad client can ruin the computations of other clients it is computing with
- **Volatility**
 - How can faults be handled without slowing or invalidating collective computation?
- **Heterogeneity**
 - User clients can be using any operating system and architecture
 - Latencies between clients range from local to global times, how can computation be distributed efficiently?

World Migrating Agent Example

Host	Location	OS/JVM	Processor
yangtze.cs.uiuc.edu	Urbana IL, USA	Solaris 2.5.1 JDK 1.1.6	Ultra 2
vulcain.ecoledoc.lip6.fr	Paris, France	Linux 2.2.5 JDK 1.2pre2	Pentium II 350Mhz
solar.isr.co.jp	Tokyo, Japan	Solaris 2.6 JDK 1.1.6	Sparc 20

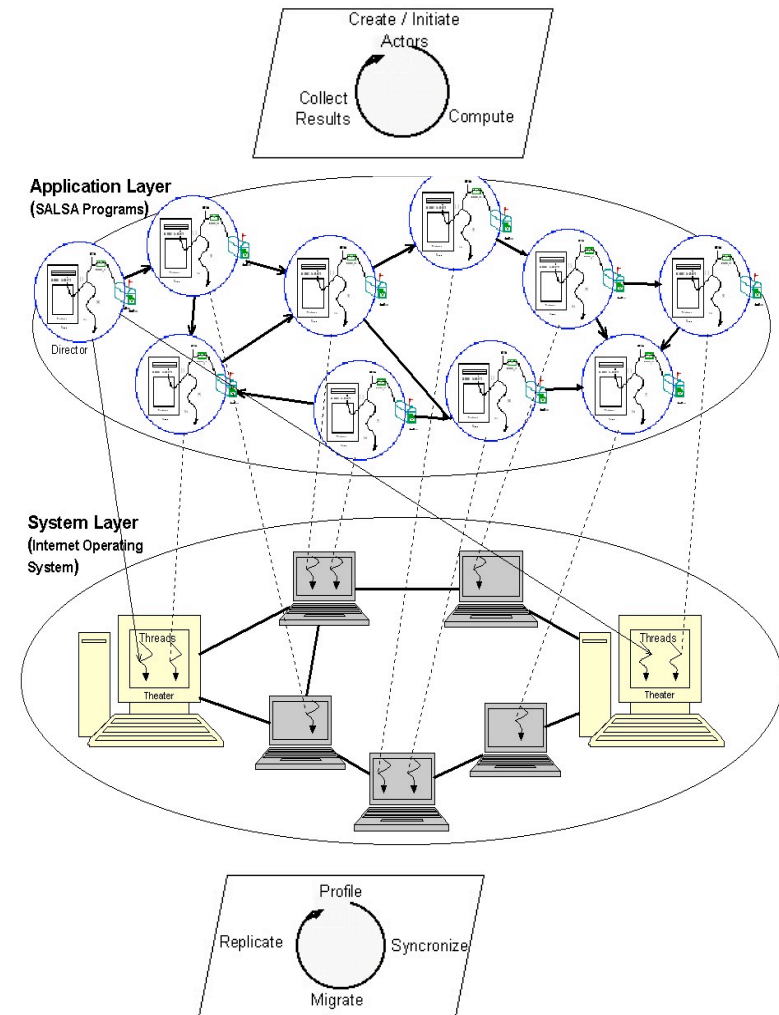
Local actor creation	386us
Local message sending	148 us
LAN message sending	30-60 ms
WAN message sending	2-3 s
LAN minimal actor migration	150-160 ms
LAN 100Kb actor migration	240-250 ms
WAN minimal actor migration	3-7 s
WAN 100Kb actor migration	25-30 s

C. Varela and G. Agha, “**Programming dynamically reconfigurable open systems with SALSA**”, *ACM SIGPLAN Notices, OOPSLA 2001*, 36(12), pp 20-34.

Middleware/IOS

- Middleware
 - A software layer between distributed applications and operating systems.
 - Alleviates application programmers from directly dealing with distribution issues
 - Heterogeneous hardware/O.S.s
 - Load balancing
 - Fault-tolerance
 - Security
 - Quality of service
- Internet Operating System (IOS)
 - A decentralized middleware framework for adaptive, scalable execution
 - Modular architecture to evaluate different distribution and reconfiguration strategies

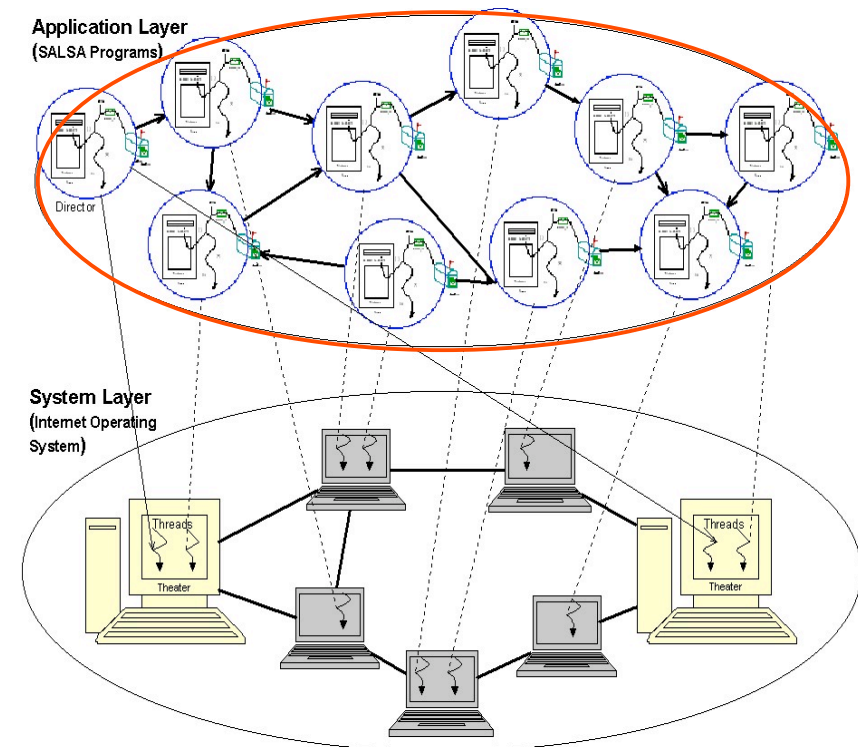
K. El Maghraoui, T. Desell, B. Szymanski, and C. Varela. **The Internet Operating System: Middleware for Adaptive Distributed Computing.** *International Journal of High Performance Computing Applications (IJHPCA)*, 2006.



Application Topology-Sensitive Work-Stealing (ATS)

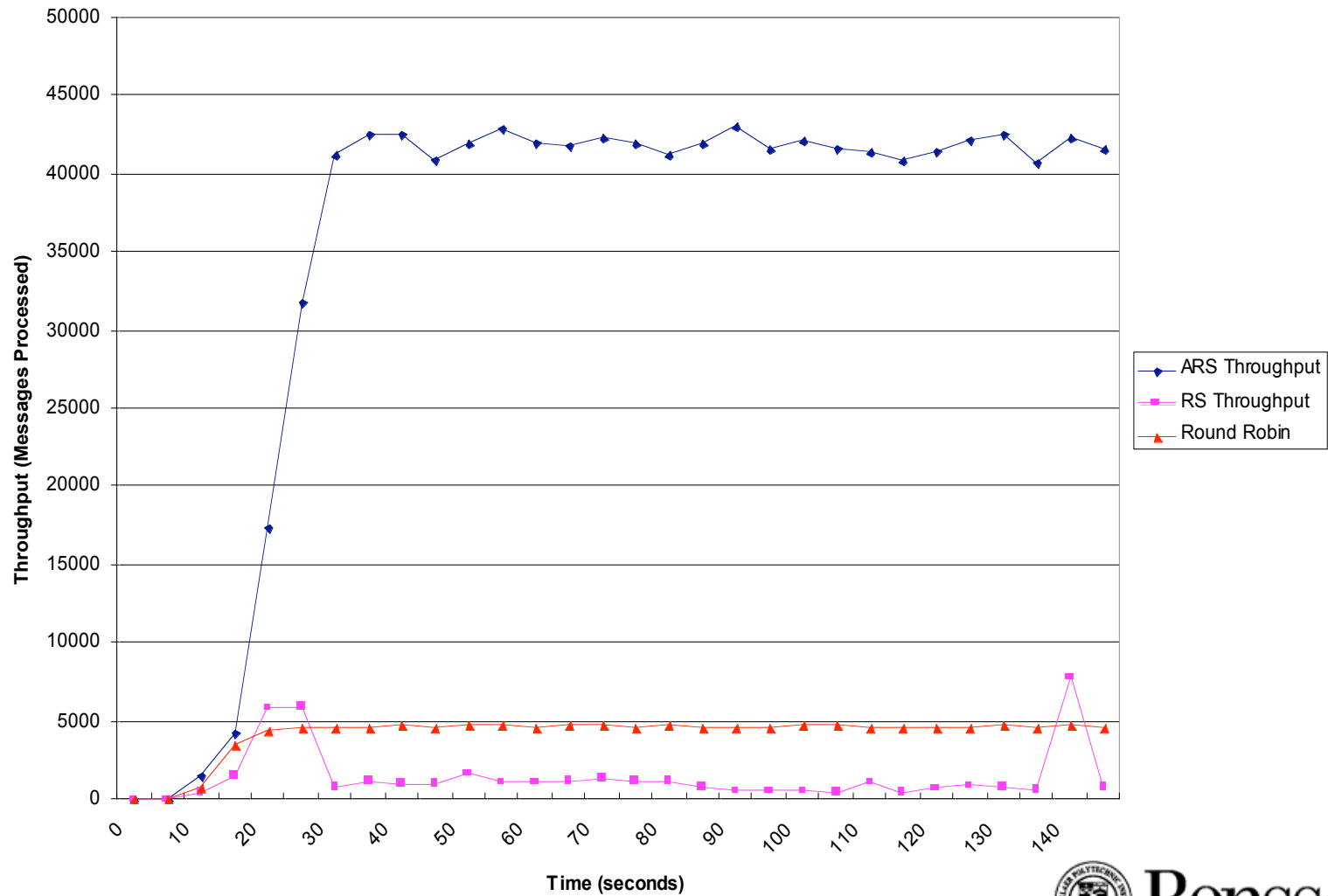
- A specialization of work stealing to collocate distributed application components that communicate frequently
- Process/actor migration to minimize high latency communication, based on
 - Location of acquaintances
 - Profiled communication history
- Tries to minimize the frequency of remote communication improving overall system throughput

T. Desell, K. El Maghraoui, and C. Varela, “**Load Balancing of Autonomous Actors over Dynamic Networks**”, *HICSS-37 Software Technology Track*, 2004.



Sparse Application Topology

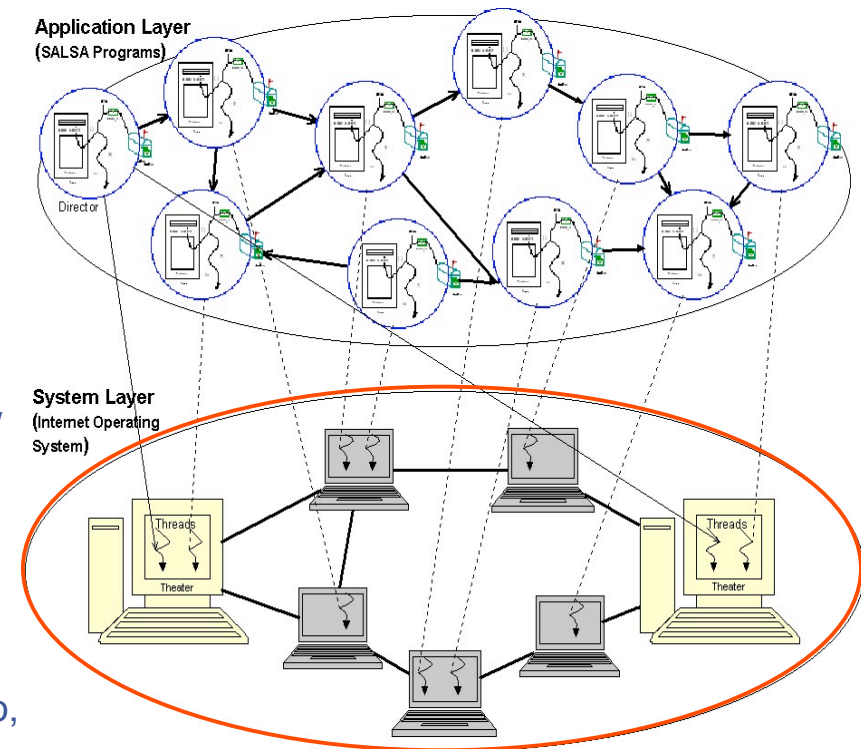
Sparse Topology



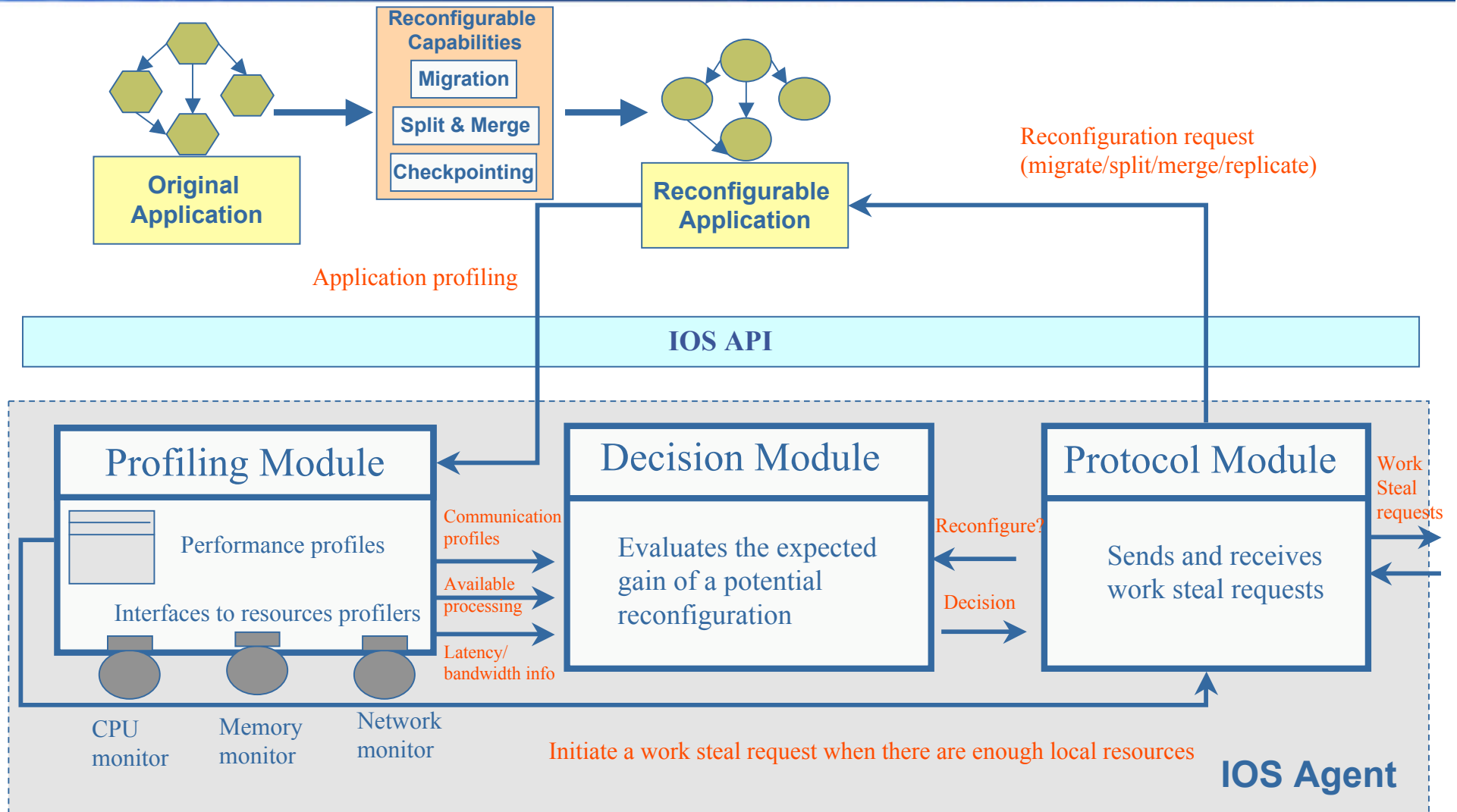
Network Topology-Sensitive Work-Stealing (NTS)

- An extension of ATS to take the network topology and performance into consideration
- Periodically profiles end-to-end network performance among peer nodes
 - Latency
 - Bandwidth
- Tries to minimize the cost of remote communication improving overall system throughput
 - Tightly coupled actors stay within reasonably low latencies/ high bandwidths
 - Loosely coupled actors can flow more freely

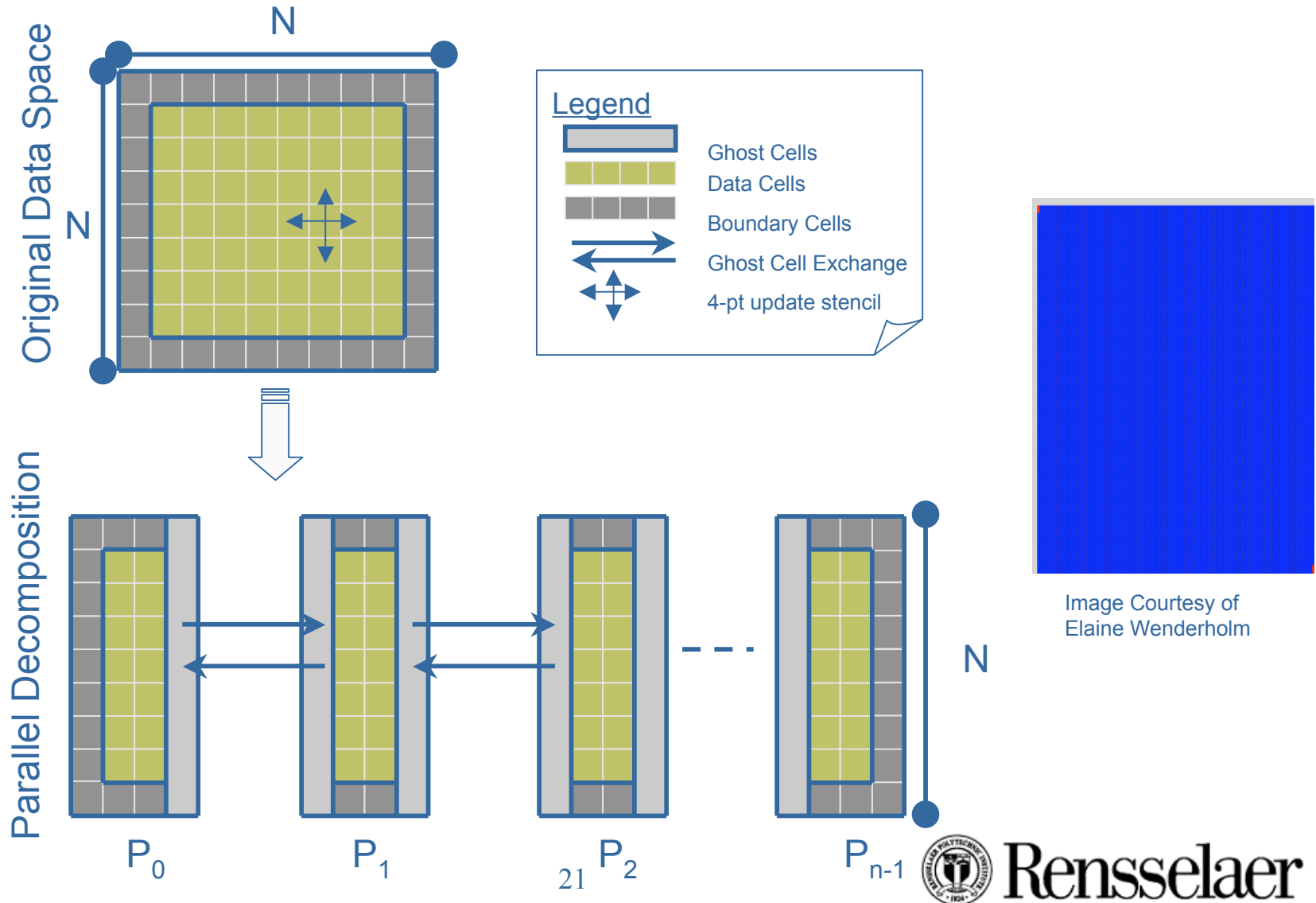
K. El Maghraoui, J. Flaherty, B. Szymanski, J. Teresco, and C. Varela, “Adaptive Computation over Dynamic and Heterogeneous Networks”, *PPAM 2003, LNCS 3019*, September, 2003.



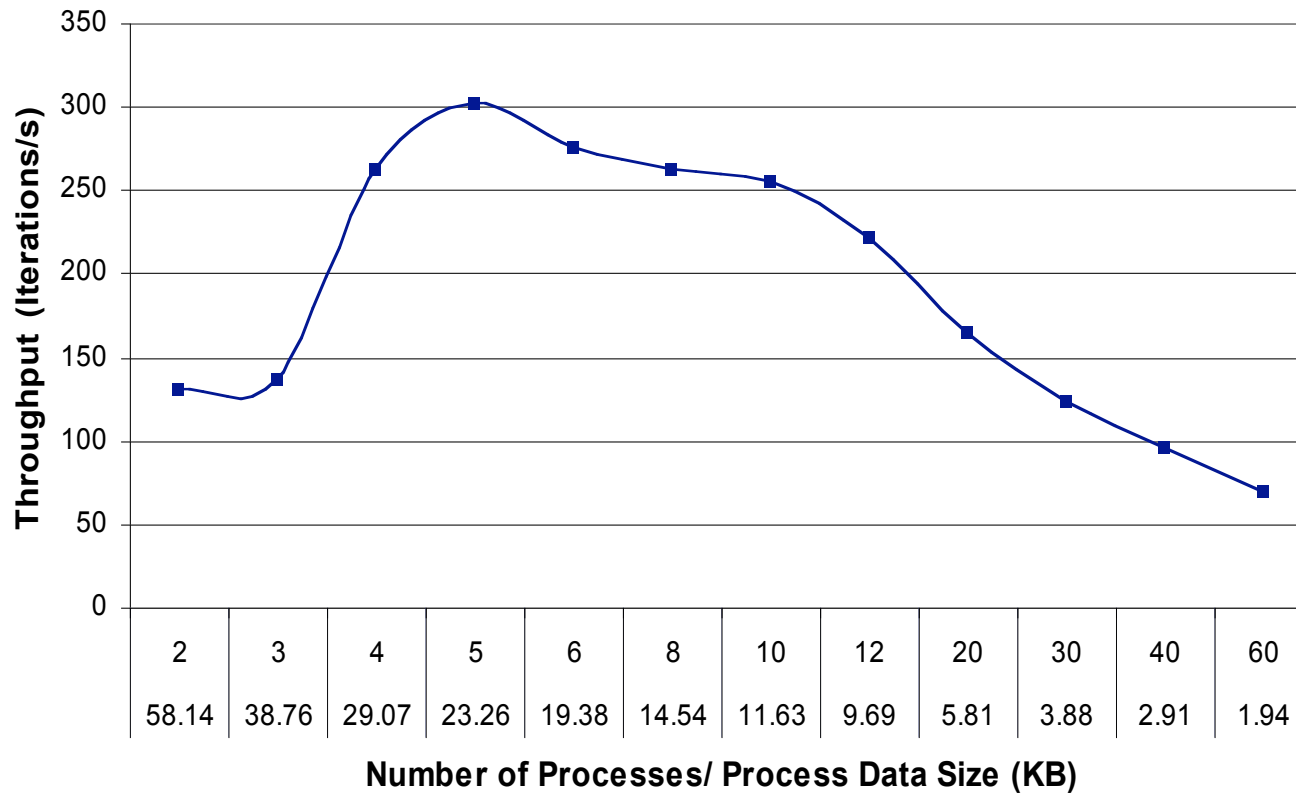
The Internet Operating System Middleware



Case Study: Parallel Decomposition of Heat Distribution

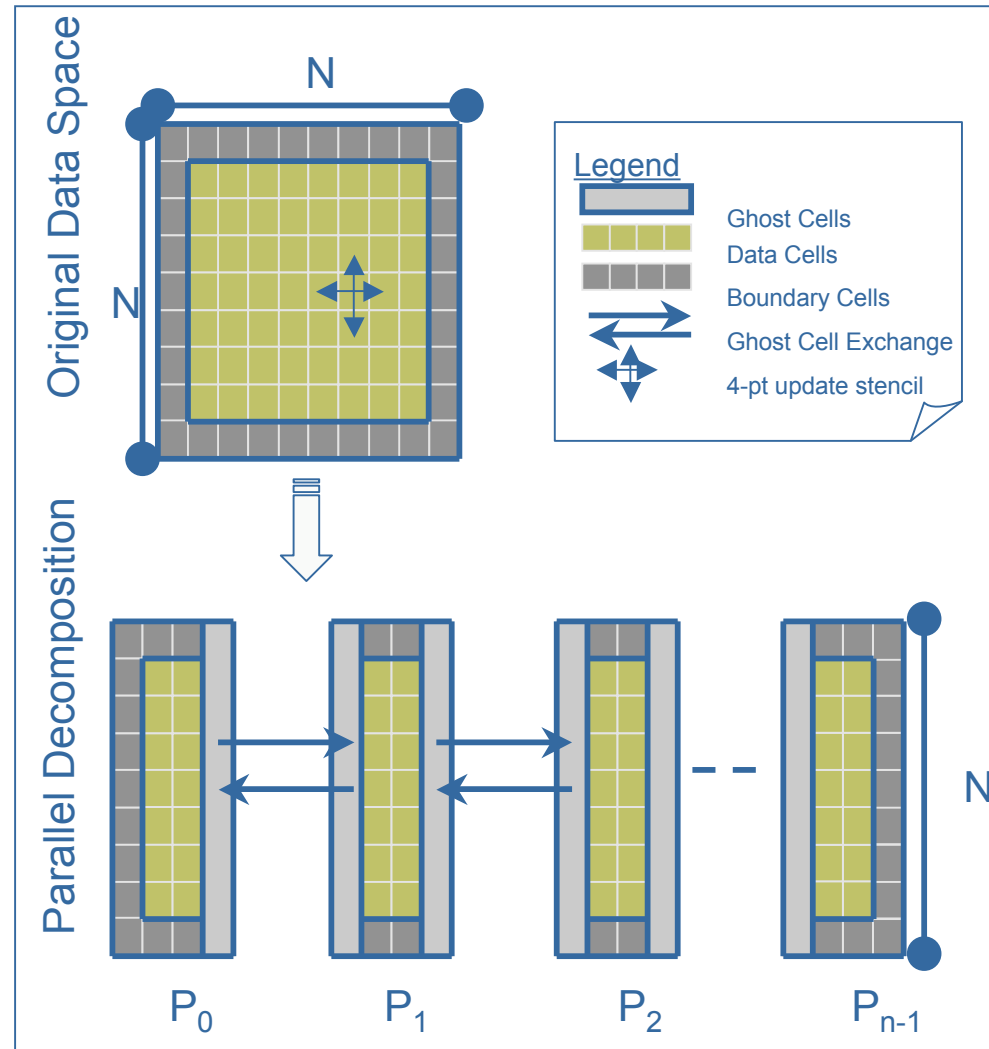
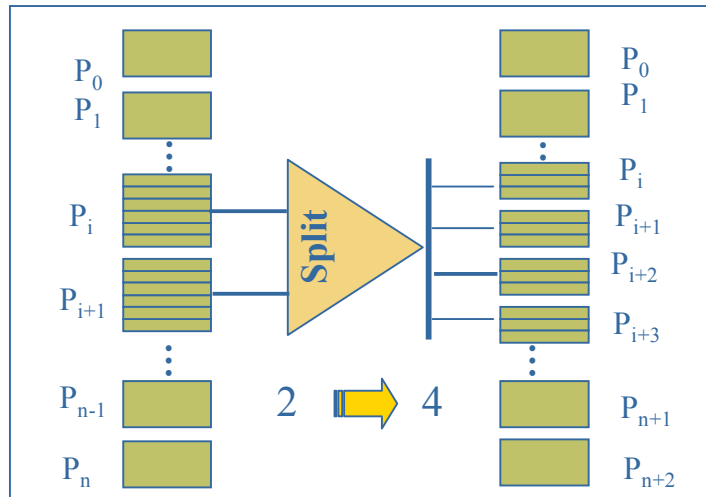
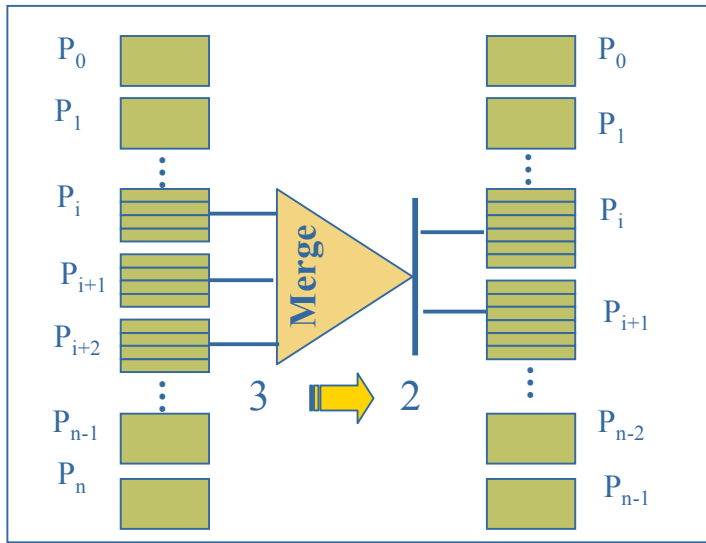


Impact of Process Granularity



Experiments on a dual-processor node (SUN Blade 1000)

Example: Split and Merge Operations



Malleable Applications

- Small process granularity
 - Generally enables better OS memory hierarchy utilization (data fits in L2 cache)
 - Incurs in higher context switching overhead
- Large process granularity
 - Requires less scheduling and context switching
 - May not use memory hierarchy efficiently (cache misses)
- **Split and merge** functionality enables application processes to dynamically change granularity
 - to improve **performance**
 - to **scale** to larger run-time environments

T. Desell, K. El Maghraoui and C. Varela, “**Malleable Applications for Scalable High Performance Computing**”, *IEEE Cluster Computing Journal*, Special Issue with Best Papers from *High Performance Distributed Computing Workshops (HPDC 2006)*, pp. 323-337, 2007.

K. El Maghraoui, T. Desell, B. Szymanski and C. Varela, “**Dynamic Malleability in MPI Applications**”, *IEEE International Symposium on Cluster Computing and the Grid (CCGrid 2007)*, pp 591–598, 2007. Best paper award nominee (top 8).

9/17/08

24

OverView

Heat Example With 5 Initial Nodes

<http://wcl.cs.rpi.edu/overview/>



Rensselaer

Security / Correctness--- Potential Directions

- **IOS uses the Java platform**
 - Can leverage JVM sandbox, bytecode authentication
- **Additional security measures**
 - Clients only accept communication from signed/trusted neighbors
 - Secure communication between applications (public/private key encryption)
- **Profiling**
 - [Client and Server Side] Detect malicious anomalous clients through stochastic result verification and redundancy
 - [Client Side] Keep a “*trust*” measure of p2p neighbors and share this knowledge, i.e., a web of trust.

Volatility --- Potential Directions

- **Stochastic Replication**

- Used for verification/correctness/trust management
- Also can be used for fault tolerance
 - Replication must be done within neighbors sharing the same shared computation

- **Algorithmic Fault-tolerance**

- Heterogeneity-tolerant failure-oblivious asynchronous algorithms
- Randomized algorithms
- Peer group checkpointing strategies

- **Fault-tolerant Programming Models**

- Lightweight process/data checkpointing/replication

J. Field and C. Varela, “**Transactors: A Programming Model for Maintaining Globally Consistent Distributed State in Unreliable Environments**”, *ACM Principles of Programming Languages (POPL’05)*, pp 195-208, 2005.

Heterogeneity --- Potential Directions

- Java based clients are designed to run on **any OS/hardware platform**
 - Is performance a significant concern? (e.g., for GPU/hybrid architectures)
- Synchronized computation requires **similar compute times among workers**
 - IOS can profile and perform dynamic reconfiguration/load balancing
 - Profiled knowledge of clients (e.g., multi-core architectures) can allow for heterogeneity-aware work scheduling and distribution
 - Group clients in similar geographical areas (low inter-client latencies) into *virtual clusters* to minimize communication overhead

Discussion

- Security of clients and servers is extremely important in volunteer grid computing – different middleware strategies can enable secure distributed computation.
- Adaptive middleware can use profiling, replication and reconfiguration (migration, malleability) to enable synchronous computing on dynamic volunteer computing grids

Merci!

Software and publications freely available at: <http://wcl.cs.rpi.edu/>

Please join MilkyWay@Home at: <http://milkyway.cs.rpi.edu/>