Streamline Computing: Cluster Integration and Effective Software

Nick Birkett

Streamline Computing Ltd

Barclays Venture Centre

Sir William Lyons Road

Coventry CV4 7EZ, UK

http://www.streamline-computing.com

General Facts

- Streamline Computing UK, founded December 2000
  - Integrators of large scale distributed memory computers.
  - Sun Sparc Solaris and PC systems.
  - Myrinet integrators.
  - Software developers for parallel tools.

General Facts

- 10 employees and gowing, with expertise in:
- Computer Science
- Scientific Computing
- Parallel Computing
- Parallel application development
- Computer system support

Company Aims

1. To profit from a major paradigm shift in High Performance Computing.
1. To establish Streamline Computing as a major provider of services AND software for distributed memory computing.

- Cluster integration support
- Application level software

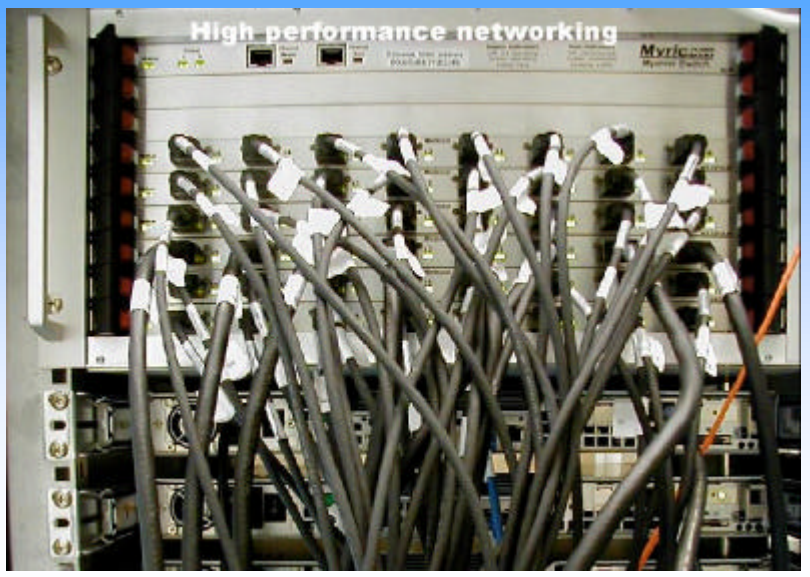

ULTRA 80
SUN BLADE 1000

Sparc Cluster 1

176 Ultra Sparc III 750MHz cpu

Score Cluster 1 - 96 PIII 866Mhz

Myrinet and Score 4.0



Founder Members

- Dr Michael Rudgyard.
- D.Phil Oxford University.
- Background in Numerical Analysis, Computational Fluid Dynamcis, Parallel Computing
- Creator of AVBP parallel computational fluid dynamics.
- Creator of COUPL and COUPL+ parallel libraries.

Founder Members (continued..)

- Dr Nick Birkett
- Phd Reading University.
- Member of Rolls Royce University Technology Centre, Oxford University 1989-1999.
- Software tools for solving large scale problems.
- System integrator for UTC.

Founder Members (continued..)

- Dr David Lecomber.
- D.Phil Oxford University.
- Background in Computer Science, parallel algorithms and analysis,
- Key developer of the COUPL+ library.

Overview

- History Of Streamline Computing and its founders 1990-2000.
- Research in Parallel Libraries.
- Beowulf systems at Oxford University.
- System Integration.
- Strategic partnerships.
- Track record.
- Software development - mid term plan

History: 1990's

- Need for high level parallel libraries.
- Low level message passing using PVM,BSP, MPI inconvenient for scientific software developer.
- Integrate common tasks such as I/O and data partitioning into library.

History: Oplus library

- One of first libraries to address needs of parallel scientific software developer.
- Developed by Prof. Mike Giles and Dr Paul Crumpton at Oxford University Computing Laboratory 1993.
- oldwww/comlab.ox.ac.uk/oucl/oxpara/parallel/oplus.html

History: Oplus library

- Code can be developed run and debugged on sequential machine.
- Same code can also run on parallel machine.
- Automatic data partitioning.
- Hides all parallel communication from developer.
- Code portability - Beowulf, Cray, SP2,SGI,..

Limitations of Oplus

- The Oplus library had some limitations:
- No parallel partitioning tool.
- Limited to explicit (order independent) calculations. Example Computational Fluid Dynamics, Electromagnetics. Time stepping methods.

COUPL+

- To address limitations of Oplus the COUPL+ library was written.
- Designed and developed by Michael Rudgyard and David Lecomber at Oxford University and at Cerfacs, France.
- Integrated parallel partitioning.
- Flexible I/O
- Parallel implicit methods.

Parallel Partitioning

- COUPL+ Distributed Partitioning.
- Parallel Hierarchical Bisection
- Parallel Inertial Bisection
- Interface smoothing

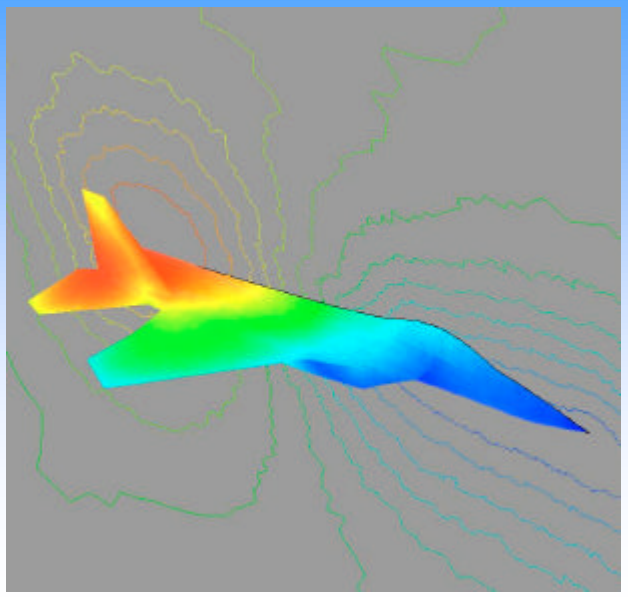
Parallel I/O model

- Basic COUPL+ model:
- - sets partitioned across processors.
- - data spans sets and is partitioned with sets.
- - automatic partitioning of pointers between sets.

Parallel partitioning

- 64 data partitions on Falcon aircraft grid
- Extremely fast algorithm -seconds to partition irregular data on parallel machine.

Parallel Visualisation



Parallel Visualisation

- Parallel Visualisation allows user to view important parts of enormous data sets.
- Use graphical workstation of modest power.
- User only interested in lines and 2 dimensional data sets (surfaces).
- User able to examine data in real time as calculation proceeds.

Parallel Loop Syntax

¢

Scalar code

```
c
c     Compute the dual volumes of a quadrilateral mesh;
c


      do n = 1, no_quads

        n1 = ielno(1,n)
        n2 = ielno(2,n)
        n3 = ielno(3,n)
        n4 = ielno(4,n)
        volcell = 0.5d0 * ( (x(2,n2) - x(2,n4))*(x(1,n1) -
x(1,n3)) -
     &                      (x(1,n2) - x(1,n4))*(x(2,n1) -
x(2,n3)) )
        voln(n1) = voln(n1) + 0.25*volcell
        voln(n2) = voln(n2) + 0.25*volcell
        voln(n3) = voln(n3) + 0.25*volcell
        voln(n4) = voln(n4) + 0.25*volcell

      end do
```

Same code in parallel

```
c
c     Compute the dual volumes of a quadrilateral mesh;
c

      do while
(kpl_par_loop(iquad_set,kpl_no_overlap,ifrom,ito,ierr))

        call kpl_access( kpl_data_access, kpl_no_access, x,
     & ielno, 1, 2, ierror)
        call kpl_access( kpl_data_access, kpl_data_access, voln,
     & ielno, 1, 1, ierror)

        do n = ifrom, ito

          n1 = ielno(1,n)
          n2 = ielno(2,n)
          n3 = ielno(3,n)
          n4 = ielno(4,n)
          volcell = 0.5d0 * ( (x(2,n2) - x(2,n4))*(x(1,n1) -
x(1,n3)) -
     &                        (x(1,n2) - x(1,n4))*(x(2,n1) -
x(2,n3)) )
          voln(n1) = voln(n1) + 0.25*volcell
          voln(n2) = voln(n2) + 0.25*volcell
          voln(n3) = voln(n3) + 0.25*volcell
          voln(n4) = voln(n4) + 0.25*volcell

        end do

      end do
```

Explanation

- kpl_par_loop is a logical function that returns the loop indices ifrom and ito, given the set identifier iquad_set.
- kpl_access is used to define how the distributed data within the loop is to be accessed. In this case the array x() is read from memory through an indirection via the connectivity array ielno(); however the array is not written back to memory. Similarly the array voln() is read from memory *and* then written back to memory through an indirection via ielno().

On the first pass of the outer loop, kpl_par_loop is true, although ito is given a value less than ifrom. The library then analyses the loop in order to decide how to copy the minimum amount of information so as to ensure that owned values of each set are up-to-date following the execution of the loop; during this or subsequent passes, the values of ito and ifrom are set accordingly, and information to be copied from remote processors is scheduled as required.

Advantages

1. The parallel code now runs on both a parallel and scalar machine.
1. The code for the main calculation of voln is virtually identical to the scalar code.
1. Programmer has only to learn some simple rules.
1. Programmer can concentrate on algorithm.

Advantages

1. Data is automatically partitioned by simple library calls.
1. User code can run using PVM, BSP, or MPI.
1. Message buffering and overlap of communication and computation optimised in library, not by programmer.

PC clusters at Oxford University Computing Laboratory

- From 1997 interest in PC clustering began, in order to supply cheap hardware to parallel software developers.
- 1997: 8 cpu PIII 450 Mhz ethernet. BSP
- 1998: 16 cpu PIII 500 Mhz ethernet. BSP (Rolls-Royce UTC research at Derby UK)
- 1999: 16 cpu PIII 800 Mhz, Score Myrinet MPI (tosca test system at Oxford University SuperComputing Centre)

1999-2000

- We had the technical expertise in following areas.
1. Scientific application development.
1. Parallel application development.
1. Parallel libraries.
1. Parallel graphics
1. Cluster integration.

2000

- Streamline Computing started.
- Founded by academics from Oxford and Warwick Universities.
- Expertise in cluster computing and parallel application development.
- Funded by venture capital and private investors.

Cluster Integration

- Company founded, October 2000.
- Started trading December 2000.
- 1st contract, 96 processor Myrinet2000 cluster.
- Raised £300,000 funding from regional development fund, venture capital and private investors.

Aims

- Turn key cluster solutions
- In partnership with SunMicrosystems and leading UK PC providers.
- Myrinet, Gigabit and ethernet solutions.
- Pre-configured software: PGI compilers,PBS, Sun GridEngine, Score.
- Support for Linux and Sparc Solaris.
- Software cluster tools.

Reasons for Starting up

- Streamline Computing started because:
- Existence of mature parallel codes: Chemistry, Bioinformatics, Fluid Dynamics,Engineering, Pharmacology.
- Distributed Parallel Computing market in UK now worth £60,000,000/year and growing.
- Few companies with expertise in UK who can provide High Performance computing.

Track Record

- Largest UltraSparc III HPC cluster in Europe (176 cpu)
- Larget UltraSparc III/Myinet2000 cluster in UK (128 cpu)
- First large Myrinet2000 cluster in UK (96 cpu)
- Clusters at 10 UK universities up to 128cpu.
- Commercial clients include Schlumberger and McLaren formula 1.

Track Record (continued..)

- To be anounced:
- - 264 cpu Myrinet2000 cluster.
- - based on 2.2GHz Intel Xeon using Intel Plumas chipset.
- - 1.2 Teraflops !!

Streamline cpus in 1st 15 months

- Bath 96 cpu
- Birmingham 52 cpu
- Durham 142 cpu
- Lancaster 176 cpu
- Leeds 48 cpu
- Oxford 224 cpu
- Warwick 122 cpu
- Woking 112 cpu

Partnerships

1. SunMicroSystems:
   - We are the main UK integrator of Sun HPC clusters.
1. Myricom:
   - We are a full distributor of Myricom Products. - over £500,000 worth installed since May 2000.
1. We are distributors of Cyclades terminal servers and Infortrend RAID storage.

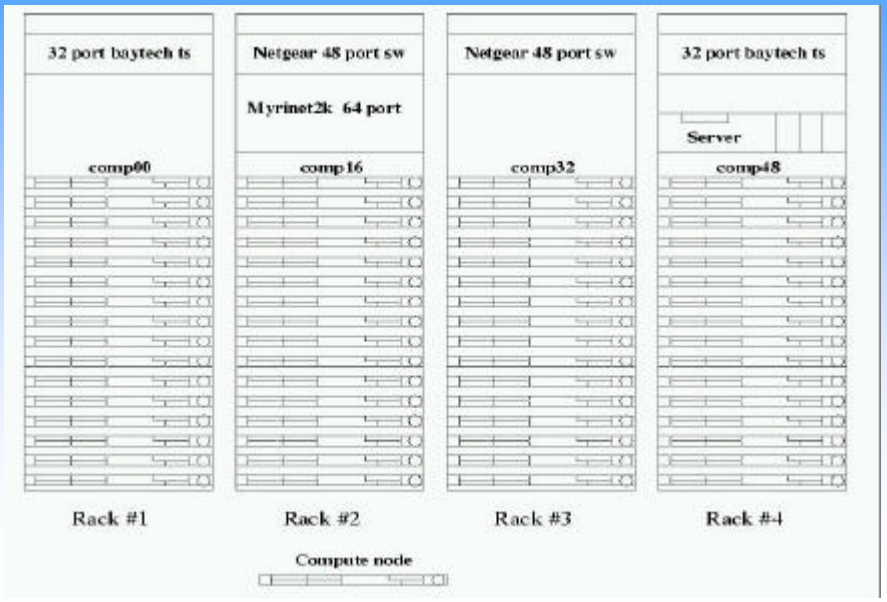
Streamline Systems

- Types of clusters used by our customers,
- Arch Network Job Management
- Linux PC Myrinet/GM PBS
- Linux PC Myrinet/SCore SCore/PBS
- Linux PC Myrinet/GM LSF
- Sun Sparc Myrinet/GM SunGridEngine

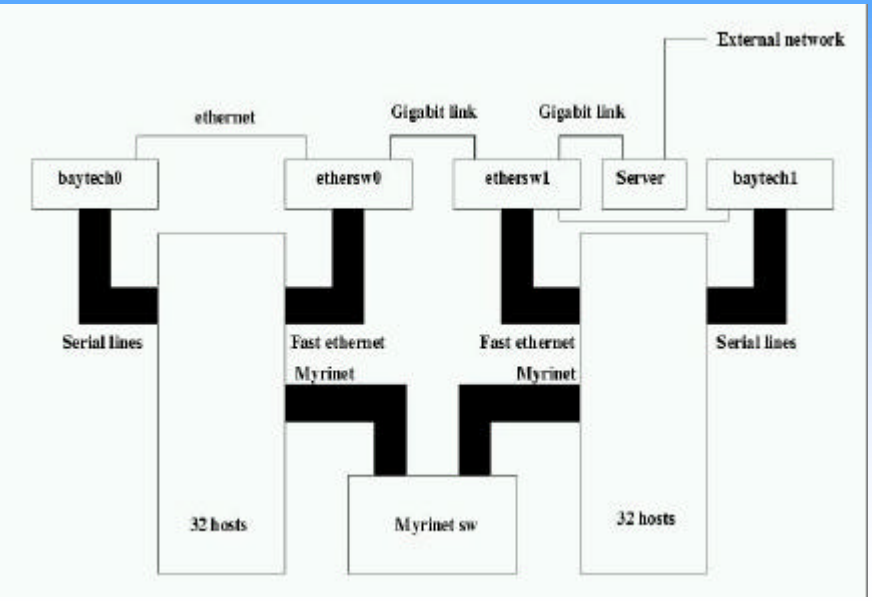
Cluster Architecture

- Clusters have following common features.
- Server node with 1Gbit uplink.
- Compute nodes
- Network switch
- Terminal server
- Myrinet switch
- Disk arrays

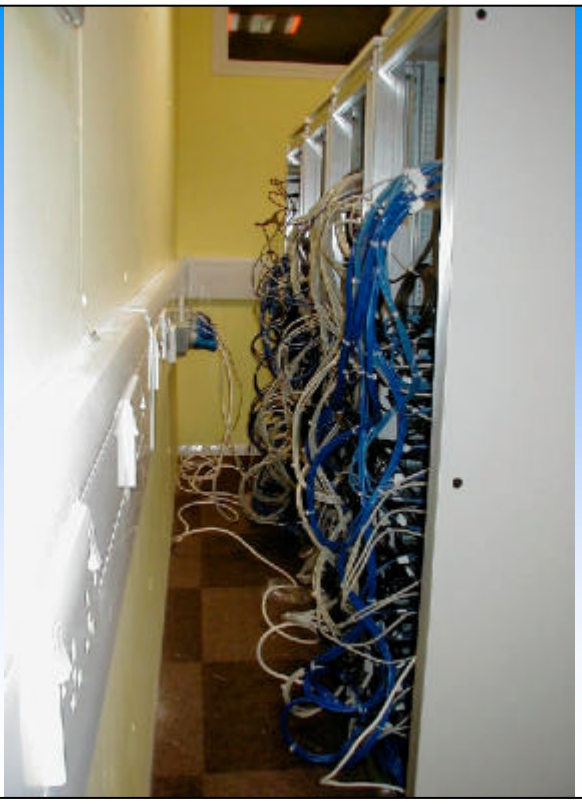
Typical Rack Mounted System



Network Topology



Under Construction



PC Cluster - components

- Typical components we use:
- SuperMicro SuperServers: www.supermicro.com
- Intel P3 and P4 Xeon systems
- AMD Athlon MP and Athlon MPX
- Myrinet 2000 2Gbit
- Cisco/Netgear switches
- Cyclades Terminal Server

Performance

- Typical performance figures for PC

System	Bandwidth MB/s	Latency us
Athlon MP 1.5	144	10
Athlon MPX 2	230	8
P3 1GHz	186	11
P4 2.2Ghz	150 -250	10

P4 cluster: 32 X 2.2GHz

Sun Clusters

- Sun is agressively pursuing the cluster market.
- Powerful 64-bit processing - large shared memory as well as distributed memory systems.
- Established solution:
- - large existing customer base.

Sun Clusters (continued)

- Viable Campus Grid solution:
  - 2,4,24, and 102-way SMP.
- Reflects diverse needs. Users can run:
  - Large memory, scalable OpenMp jobs.
  - Mildly scalable OpenMp jobs
  - Scalable MPI jobs.
- Appropriate Software stack:
  - GridEngine

Sun Clusters (continued)

- Campus grid solutions:
  - Full range of Sun storage and backup solutions.
  - Embedded Linux clusters.

Parallel Applications

- Some parallel codes used by our customers:
  - DLPOLY (Chemistry)
  - ADF (Chemistry)
  - MOLPRO (Chemistry)
  - CHARMM (Chemistry)
  - GROMACS (Chemistry)
  - NWCHEM (Chemistry)

Parallel Applications

- Some parallel codes used by our customers:
  - CASTEP (Materials)
  - STARCD (Fluid Dynamics)
  - HYDRA (Fluid Dynamics)
  - FLUENT (Fluid Dynamics)
  - HYDRA (Astro-physics)
  - ECLIPSE (Oil well reservoir simulation)

Trends in Computing

- Fast growth in distributed computing market fuelled by:
- Cheap and increasingly powerful hardware - Intel P4 Xeon, Athlon, Intel IA64. Myrinet.
- Availability of parallel applications.
- Effective cluster management systems: LSF,Score,PBS,SunGridEngine.

Comparison - top 4 (taken from top500 Nov 2001)

Country	Count	Share	cpu
USA	230	46%	109681
Germany	59	12%	14734
Japan	57	11%	11896
UK	34	7%	7038

Strategic development

- Procurements will often not be made on price.
- Streamline aims to provide significant ?added-value?
- Developing a strong technical support team.
- Developing software tools and environments.
- Strategic aim is to develop a comprehensive environment for distributed computing

Summary of our business model

- There are 2 sides to our business:
- System Integration.
- Software development.

System Integration

- We do NOT buy and sell PC hardware but work with both Sun resellers and PC manufacturers.
- We are system integrators for Myrinet.
- We are not tied down to one hardware supplier.
- We coordinate and build systems, with racks,shelves cabling, network and software.

Working with Sun and PC suppliers

Software Development

- We are developing the following products:
- Model simplification - tools for visualising very large data sets.
- Parallel development tools. Our DDT Distributed Debugging Tool is approaching beta release.
- Next generation parallel libraries for distributed systems.

Software

- Parallel Programming (the past 10-12 years)
- 10-12 years ago we used shared memory directives
- Now we use OpenMP? .
- Message passing:
- 10 years ago we used PVM
- Now we use MPI? .
- Have we really progressed the state of the art ?

Future

- The future of distributed computing ?
- Enormous investment in middleware?
- But middleware has to interact with real applications?
- We also need capabilities at the application level !

Next Generation Library ?

- A simple programming environment for scaleable parallel code?
- Code that is easy to develop and debug
- Easy to maintain and extend
- Works on clusters and SMP boxes
- Key stumbling blocks
- Require shallow learning curve (avoid big libraries !)
- Automatic data-placement, effective i/o

Next Generation (continued)

- Productivity capabilities
- Self-optimising code (run time !)
- Grid-enabling tools (submission wrappers)
- Efficient parallel I/O model
- Check-pointing
- Parallel visualisation capabilities
- Computational steering capabilities
- Use of optimised libraries

System manager's wish list

- Fault-tolerance
- Job migration and checkpointing
- Throughput optimisation
- ?Intelligent?application should communicate with real-time scheduler
- Real time job contraction / expansion
- Application ?cookies?will report past performance
- Could we triple throughput ?

Streamline's vision

- An integrated approach is the only way to achieve much of the above? .
- Our aim is to make some tentative first steps
  - Advanced talks on £1-1.5M investment.
- Prototypes at the heart of real applications already exist for:
  - 3D unstructured mesh, multigrid FE applications
  - Maps to some other scientific problems

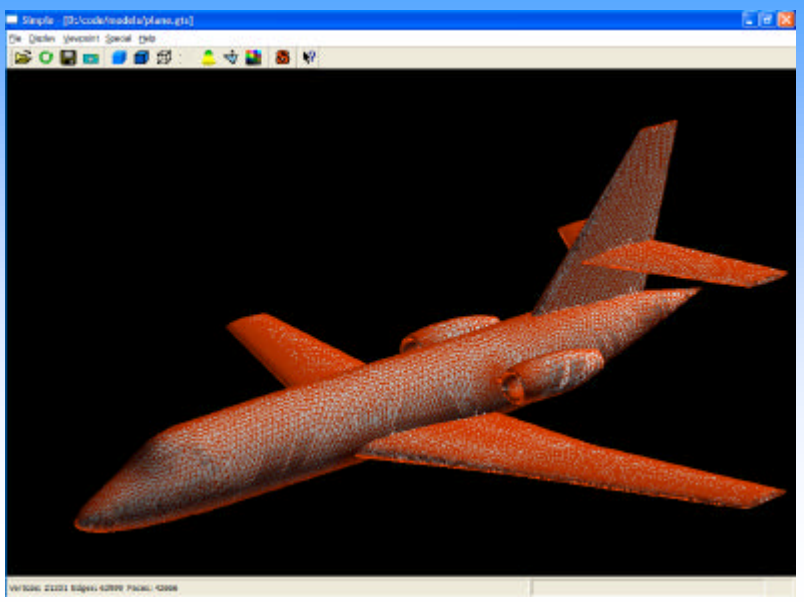
Software spin offs

- Parallel debugger and performance tools.
- Parallel visualisation software.
- Parallel partitioning capability.
- Distributed model simplification and compression (for viewing on a workstation, while minimising bandwidth?).
- Medium term aim to embed this in commercial software.

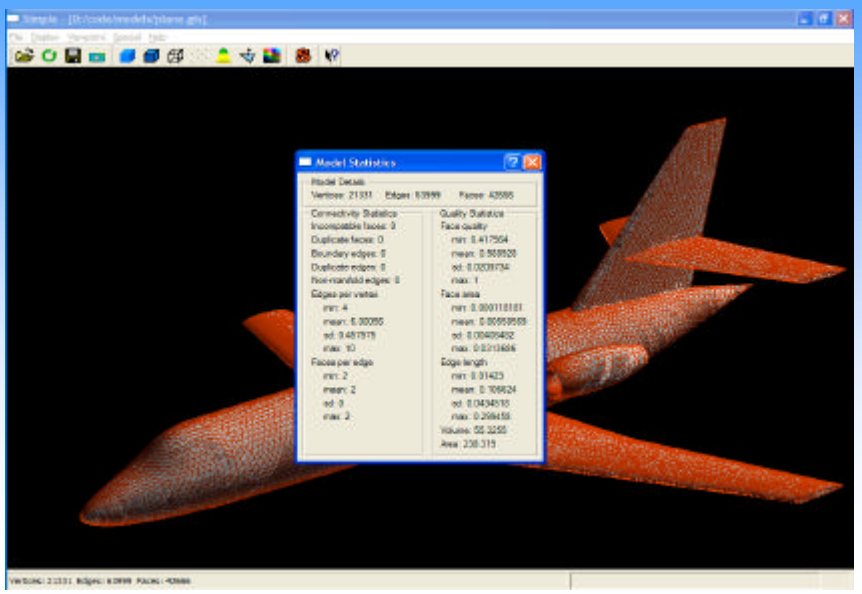
Model Simplification

- Modern simulation software on parallel computers can generate huge amounts of data.
- We make sofware tools for postprocessing output data.
- Vastly reduces data size.
- Enables visualisation on modest workstation.

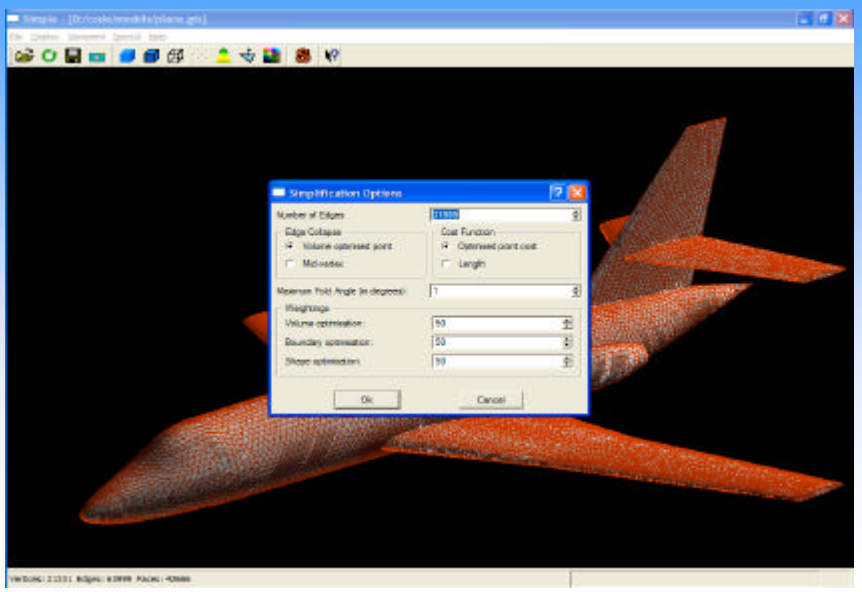
Aircraft mesh



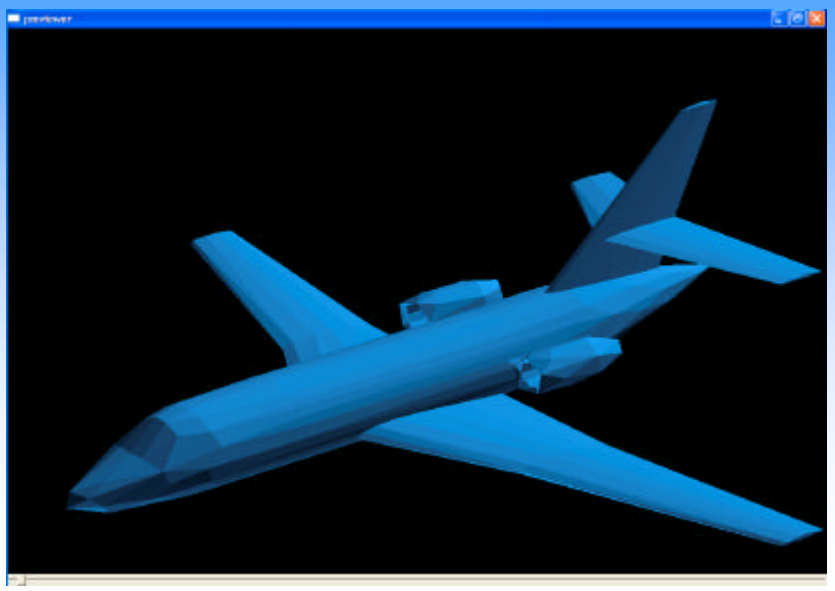
Viewing statistics



Simplifying the model



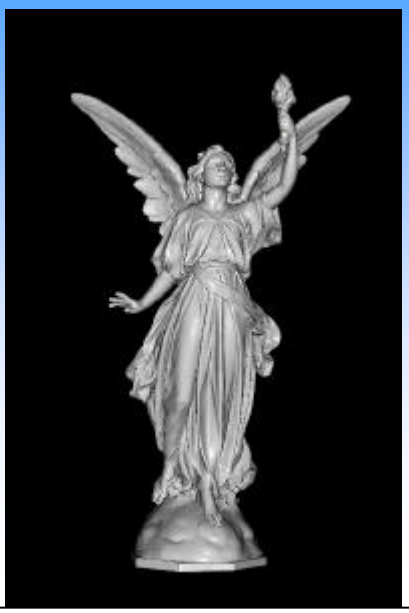
Low Resolution Reconstruction



Model Simplification

- Uses the memoryless model described by Turk and Lindstrom.
- Works on edges, triangulated meshes.
- Can run in parallel.
- This software is part of our continuing development for Parallel Graphics Libraries.

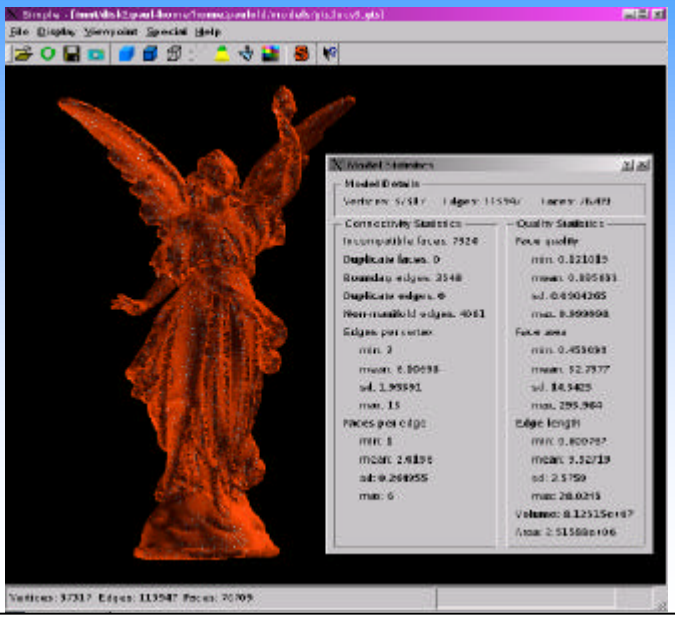
Example Using Laser Scan Image



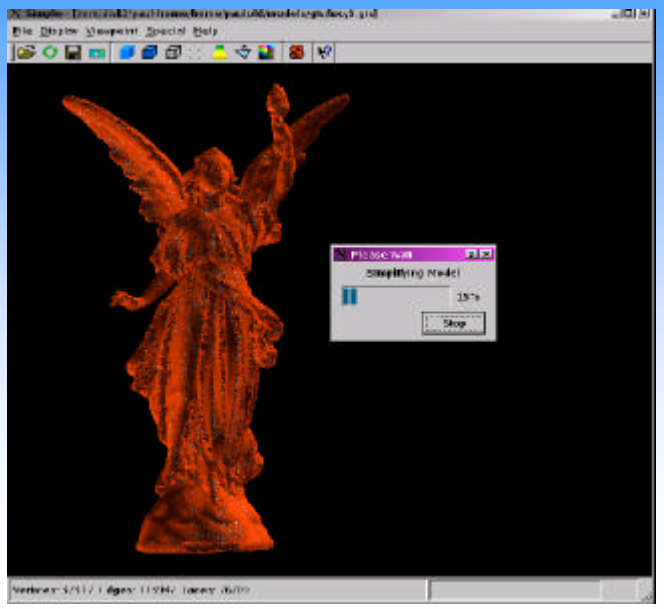
Model Simplification

- Original scan takes 0.5 Gbyte data
- Simplified model take 3Mbye data

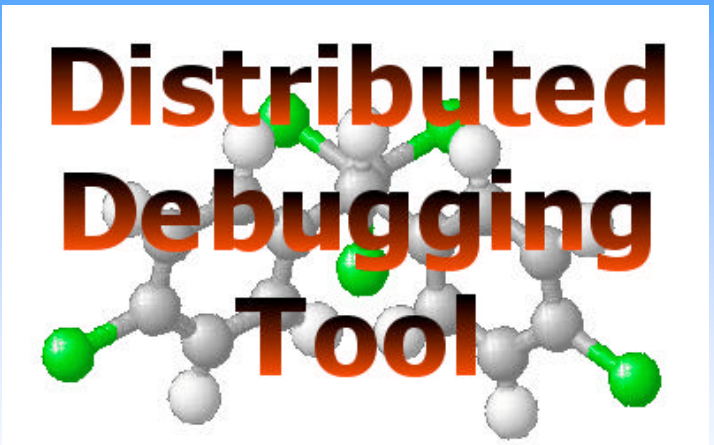
Simplified model



Simplified Model

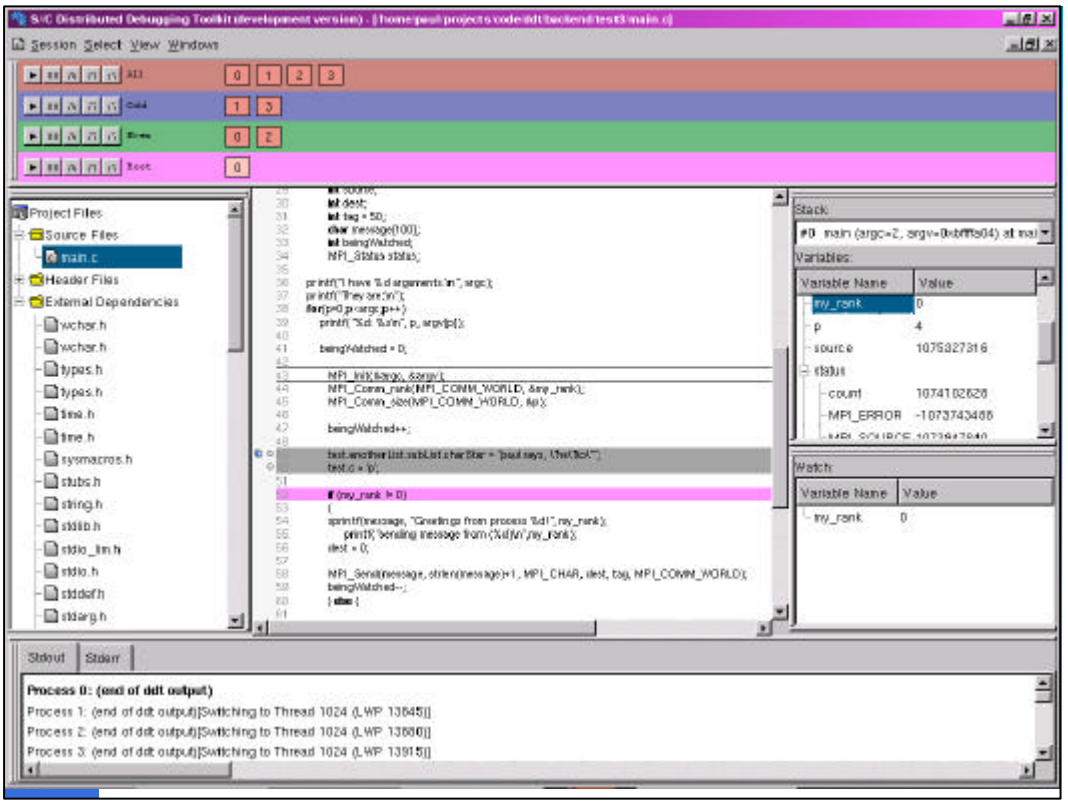


Introducing DDT



DDT

- Graphical front end user interface
- Uses QT graphics
- "Drag and drop"
- Process groups
- Break points
- Watch variables
- Complete MPI debugger.



DDT

- DDT will be released as beta test in next few weeks.
- Hooks in debugger will enable it to interface with most commercial compilers eventually.
- Esimated cost of final version: $1000 for 16 cpu, rising to $5000 unlimited at single site.
- We welcome beta testers to use on large codes.

Future

- We expect the cluster computing market to grow at an ever increasing rate.
- Cluster integration business will continue to grow over the next 2-3 years.
- We see the importance of developing effective cluster software and bringing this to market.

Summary

- The Cluster HPC market is an exciting and new business.
- To make the most of Clusters we need the right hardware AND software.
- Streamline Computing is a new company which seeks to make the most of clusters.
- Streamline Computing seeks stategic alliances to do this.

Conclusion

- We are an important player in the UK HPC market.
- We?d like to be a global player in HPC software? .

Acknowledgements

- We would like to thank the following:
- RIKEN Japan. Many thanks to Ryutaro Himeno, RIKEN. Without the generous funding from RIKEN this visit would not have been possible.
- PCCLUSTER consortium. Many thanks and in particular Yutaka Ishikawa for organising this event. We also appreciate the help and support of all the SCORE team over several years.