Supercomputing on a Shoestring: Experience with the Monash PPME Pentium Cluster

by Carlo Kopp

Computer Science & Software Engineering

Monash University, Australia

© 1999, Carlo Kopp

Introduction:

- Clusters of low cost commodity machines are a viable alternative for many high performance computing applications.
- In 1998 Monash CSSE commissioned the PPME Pentium Cluster for the purpose of running large parametric simulations.
- This presentation discusses background issues, and some of the experience acquired in porting and running large simulations on this system.

© 1999, Carlo Kopp

Supercomputer Limitations:

- Very expensive, unaffordable for small or medium sized research projects.
- High performance usually results from specialised vector processing hardware.
- Performance on jobs which do not vectorise efficiently can be mediocre and uncompetitive against cheaper machines.
- What are the alternatives ?

© 1999, Carlo Kopp

COW, NOW, PoPC:

- COW - Cluster of Workstations
- NOW - Network of Workstations
- PoPC - Pile of PCs
- Utilise inexpensive commodity processor and high speed switch hardware.
- Aggregate CPU cycles and RAM size can be competitive against supercomputer category machines.

© 1999, Carlo Kopp

How to Parallelise ?

- Parallelising/vectorising compilers.
- Code the application from the outset for distributed processing on COW/NOW/POPC.
- Utilise parametric processing tools.

© 1999, Carlo Kopp

Parametric Computing:

- Suitable for problems where a single program must be executed repeatedly with different initial conditions.
- Each CPU in the COW/NOW/PoPC runs an instance of the program with a different set of initial conditions.
- Sequential execution on a single very fast CPU replaced by parallel execution on many not so fast CPUs.

© 1999, Carlo Kopp

Activetools Clustor:

- Developed from the Nimrod parametric processing tool.
- Commercialised in the United States.
- Ported to Intel/Linux, SPARC/Solaris, MIPS/Irix, PowerPC/AIX, HP-PA/HPUX, Alpha/DU, Intel/NT.
- Root node emulates a gaggle of "robot users" who each execute a job on a client node, each job has unique runtime parameters.
- Transparently parametrises command line or file arguments at job runtime.

© 1999, Carlo Kopp

Monash PPME Cluster:

- Parallel Parametric Modelling Engine
- Initially set up August 1998 at Clayton CSSE
- 1 x dual 333 MHz P-II root node (Linux), 10 x dual 333 MHz P-II client nodes (Linux/NT).
- Upgraded early 1999 with 4 x dual 500 MHz P-III client nodes.
- Linked to Caulfield cluster with 16 x dual 350 MHz P-III client nodes.
- Currently 60 x Pentium CPUs, 5.8 GB RAM, 180 GB Disk, 2 x 100 Mbit/s switches.

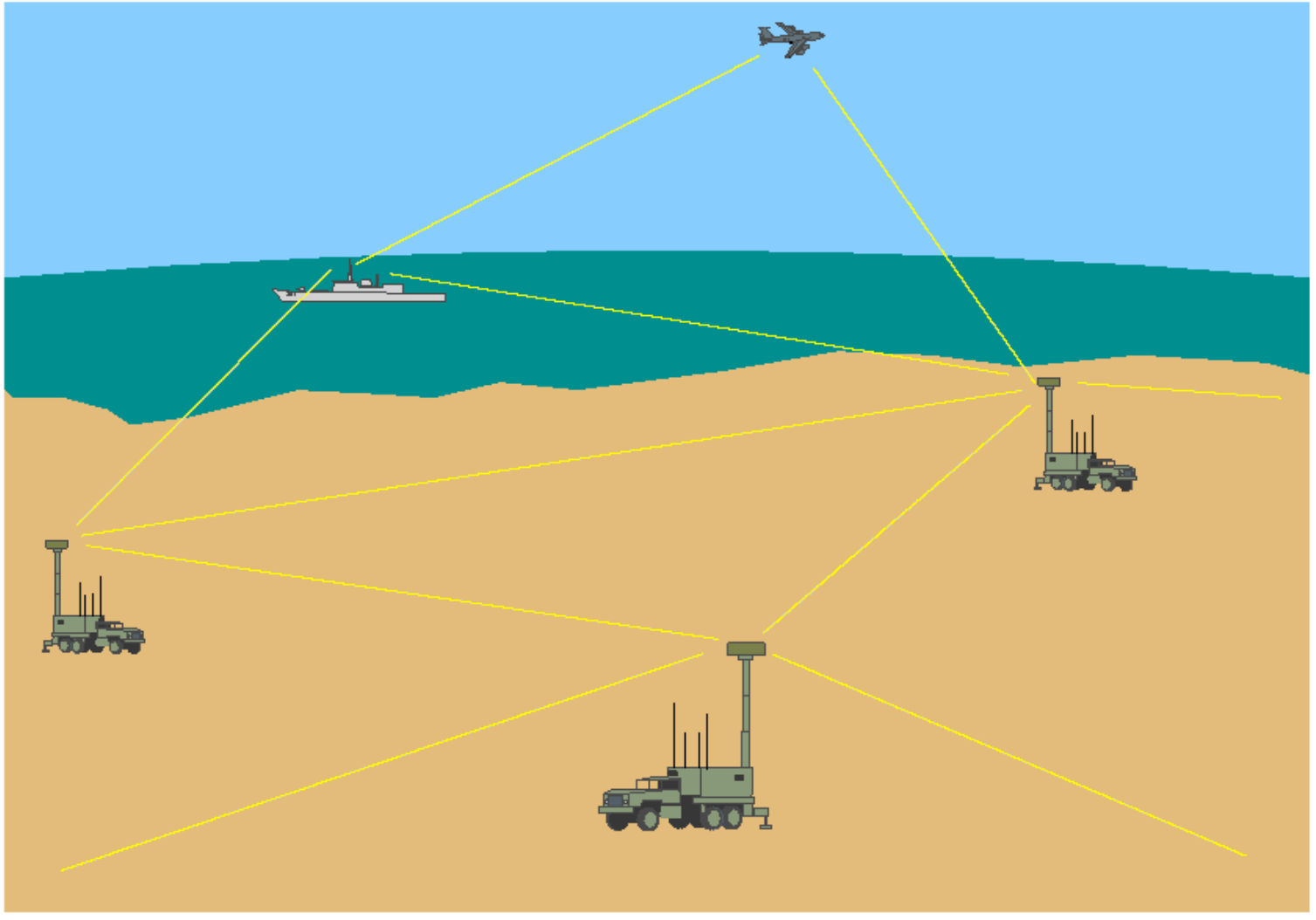
© 1999, Carlo Kopp

© 1999, Carlo Kopp

Simulation Problem:

- Ad Hoc Mobile Networks with thousands of nodes.
- Need to explore acheivable network performance at different microwave frequencies, under different weather conditions.
- Must consider local tropospheric refraction and propagation impairments.
- Must analyse global network performance.

© 1999, Carlo Kopp



© 1999, Carlo Kopp

Problem Issues:

- Total number of simulations ~ 1000
- Each simulation has up to 5 initial conditions.
- Each simulation job ~8-72 hrs on P-II/III
- Managing simulation activity is difficult.
- Sorting results and postprocessing difficult.
- Classical parametric computing problem.

© 1999, Carlo Kopp

Adaptation to Clustor:

- Disable X11 activity display for client nodes.
- Port code form FreeBSD/Irix versions to Linux.
- Regression test and validate simulation results on Linux.
- Incorporate journalling and restart facilities.
- Write and debug Clustor "Plan Scripts".
- Set up directories for results.
- Map out parameter space for simulations.
- Test runs to debug and validate operation.

© 1999, Carlo Kopp

Hardware Issues:

- Root and Client node hardware reliability must be very high.
- Client node swap space sizing.
- Client node memory sizing to preclude swapping.
- Hardware must be stable.
- Network must be stable.

© 1999, Carlo Kopp

Operating System Issues:

- Red Hat Linux reliability and stability not sufficient for the root node, although adequate for client nodes.
- NFS and TCP/IP stack integrity was the biggest source of difficulty, proved to be unreliable under heavy I/O load.
- The root node's operating system is the single point of failure for the whole cluster and it must have exceptional reliability and integrity.
- TurboLinux more suitable for root node OS.

© 1999, Carlo Kopp

Managing Multiuser Loads:

- Clustor imposes limits on the number of jobs per user per client node.
- Clustor imposes limits on the total number of jobs per client node.
- With many contending users, large jobs tended to displace small jobs, penalising light users -> “Cluster Hogging”.
- Monash developed a background scheduler scheme using the Unix nice facility.
- Jobs are periodically reniced by cumulative run time to favour shorter running jobs.

© 1999, Carlo Kopp

Simulation Strategy:

- Split the parameter space into blocks to avoid runs which last longer than 1-2 weeks.
- Use cluster toolset to regenerate runs after crashes.
- Ensure results are properly organised to avoid redundant job execution.
- Maintain a status chart to follow progress.

© 1999, Carlo Kopp

Simulation Results:

- Highly successful project.
- 99% of parameter space covered by simulations.
- PPME cluster allowed much more ambitious simulation effort than originally planned 2 years ago.

© 1999, Carlo Kopp

Summary:

- Monash PPME == Parametric Supercomputer.
- Implementation Cost ~ A$100K.
- Commodity Pentium II/III CPU hardware.
- Public Domain Operating System.
- Commodity 100 Base T switches.

© 1999, Carlo Kopp