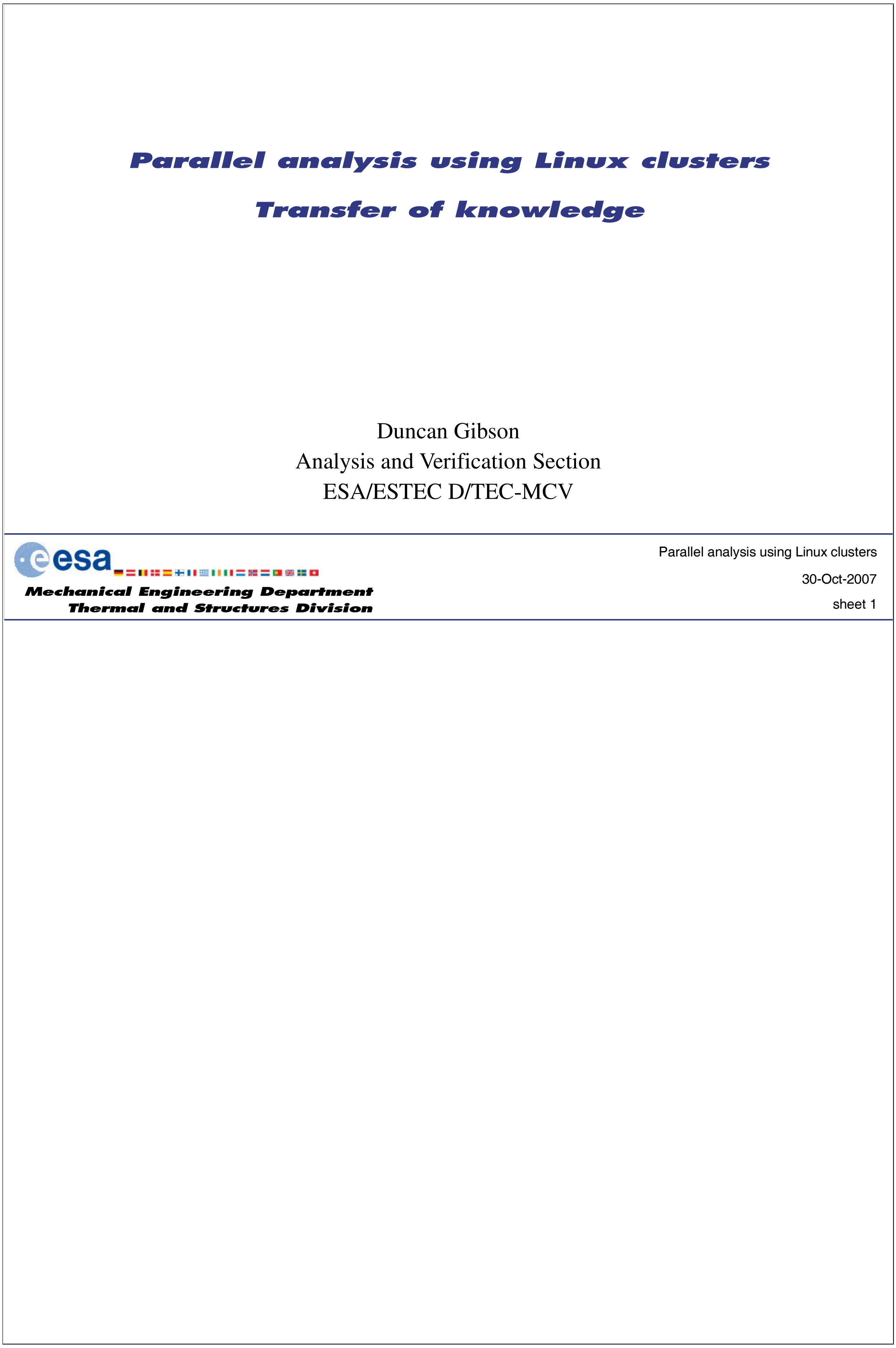
Appendix B

Parallel thermal analysis with Linux clusters

Duncan Gibson
(ESA/ESTEC, The Netherlands)



Parallel analysis using Linux clusters

Transfer of knowledge

Duncan Gibson
Analysis and Verification Section
ESA/ESTEC D/TEC-MCV

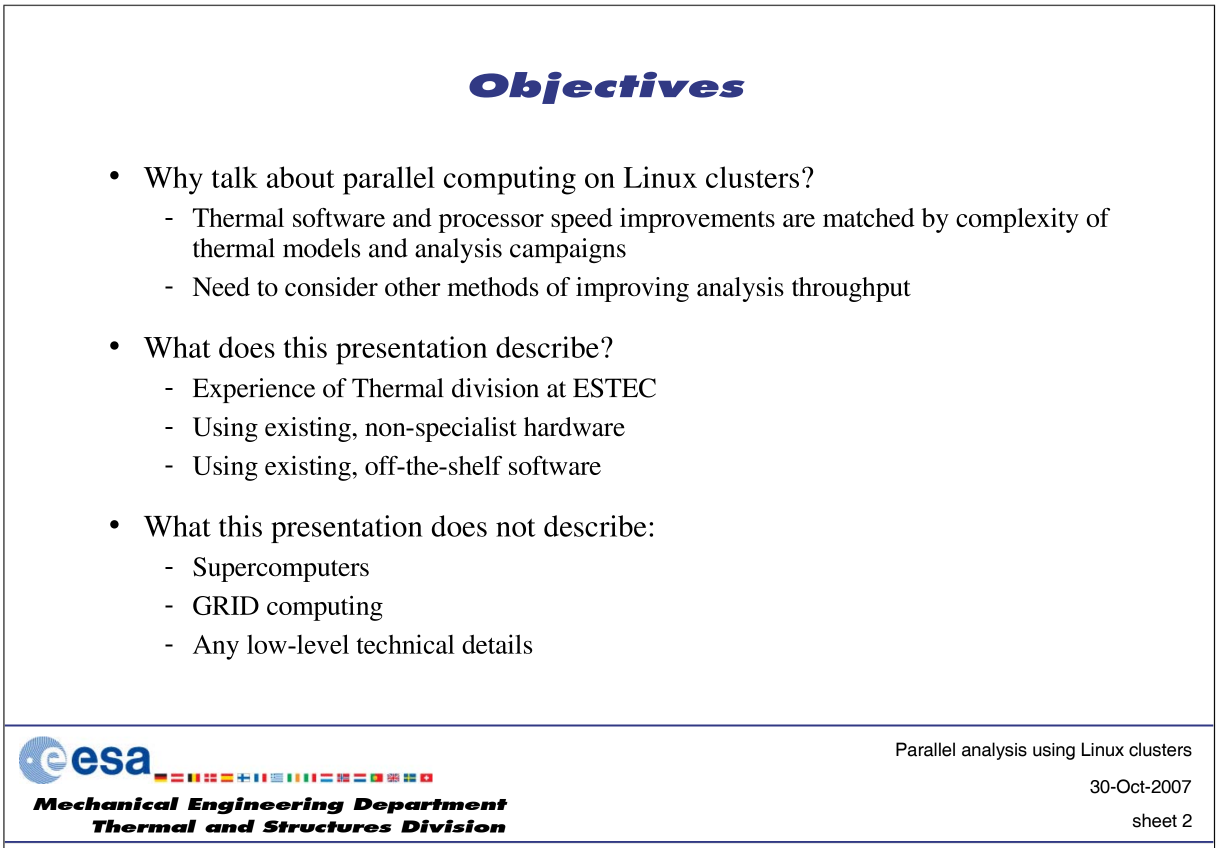
esa

Mechanical Engineering Department
Thermal and Structures Division

Parallel analysis using Linux clusters

30-Oct-2007

sheet 1



Objectives

- Why talk about parallel computing on Linux clusters?
  - Thermal software and processor speed improvements are matched by complexity of thermal models and analysis campaigns
  - Need to consider other methods of improving analysis throughput
- What does this presentation describe?
  - Experience of Thermal division at ESTEC
  - Using existing, non-specialist hardware
  - Using existing, off-the-shelf software
- What this presentation does not describe:
  - Supercomputers
  - GRID computing
  - Any low-level technical details

esa

Mechanical Engineering Department
Thermal and Structures Division

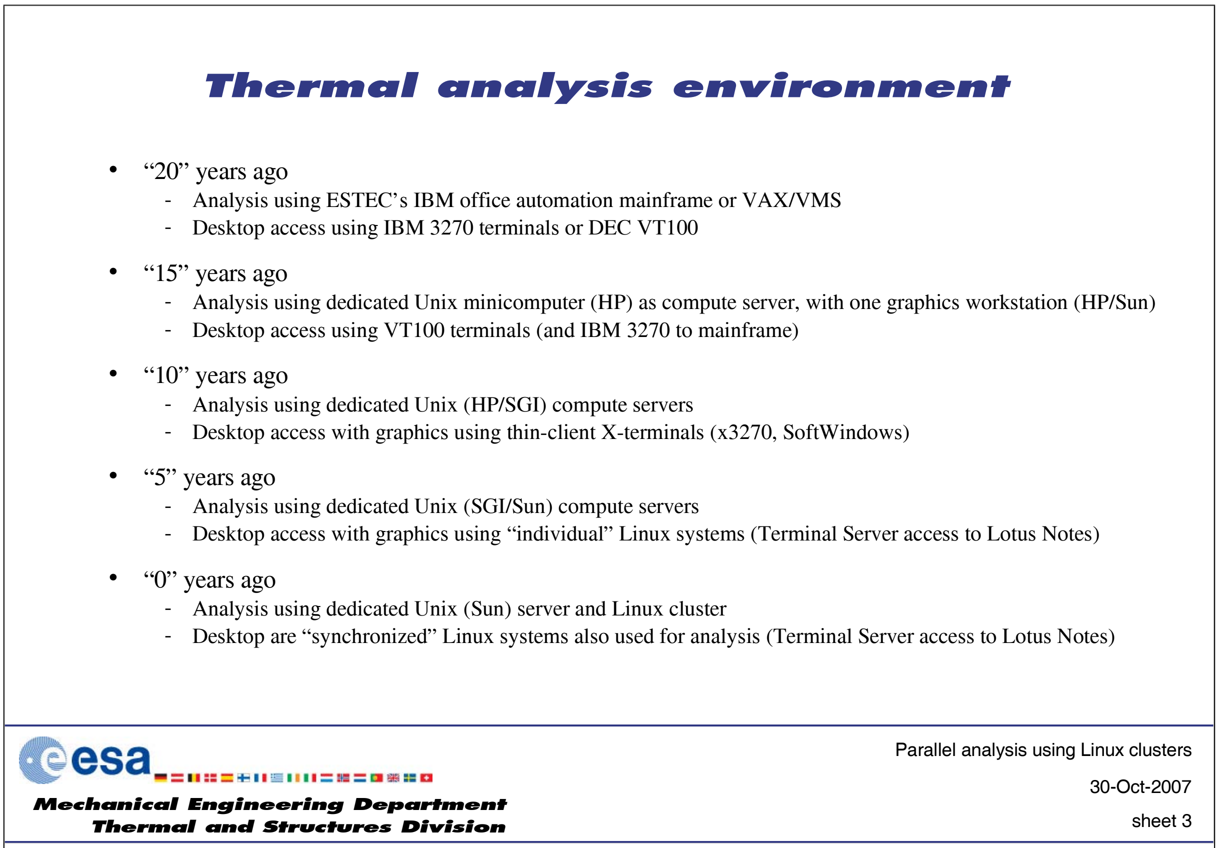
Parallel analysis using Linux clusters
30-Oct-2007
sheet 2

Previous workshops have discussed model reduction techniques to speed up analysis, but nobody seems to be using them. In fact the trend is the opposite with bigger models, perhaps from CAD systems, that soak up any software or processor speed improvements. Parallel computing provides another approach to improving analysis throughput. To support it properly, thermal managers need to consider their infrastructure, and tool developers need to design for parallel use.

ESTEC has maybe a little more freedom to explore some areas because we don’t have the same drivers as industry. This is one area in which we have some experience, and it is worth sharing that experience. We have just scratched the surface, but what we have done might also be possible for everyone, without major capital investment, today.

Discussion aimed at typical thermal departments using local hardware. Not aimed at supercomputers for weather forecasting or nuclear reactor simulations, nor at remote execution or grid computing used in some climate studies and SETI@home

This presentation is just an overview to get people thinking. Don’t want to go into any really detailed technical descriptions because everyone’s environment will be different.



Thermal analysis environment

- “20” years ago
  - Analysis using ESTEC’s IBM office automation mainframe or VAX/VMS
  - Desktop access using IBM 3270 terminals or DEC VT100
- “15” years ago
  - Analysis using dedicated Unix minicomputer (HP) as compute server, with one graphics workstation (HP/Sun)
  - Desktop access using VT100 terminals (and IBM 3270 to mainframe)
- “10” years ago
  - Analysis using dedicated Unix (HP/SGI) compute servers
  - Desktop access with graphics using thin-client X-terminals (x3270, SoftWindows)
- “5” years ago
  - Analysis using dedicated Unix (SGI/Sun) compute servers
  - Desktop access with graphics using “individual” Linux systems (Terminal Server access to Lotus Notes)
- “0” years ago
  - Analysis using dedicated Unix (Sun) server and Linux cluster
  - Desktop are “synchronized” Linux systems also used for analysis (Terminal Server access to Lotus Notes)

esa

Mechanical Engineering Department
Thermal and Structures Division

Parallel analysis using Linux clusters
30-Oct-2007
sheet 3

This is a very rough guide to the evolution of the thermal analysis environment available to the engineers at ESTEC. The timeline and system details are approximate because new systems were phased in, and old systems were phased out, so there was a much more gradual evolution or metamorphosis than appears above.

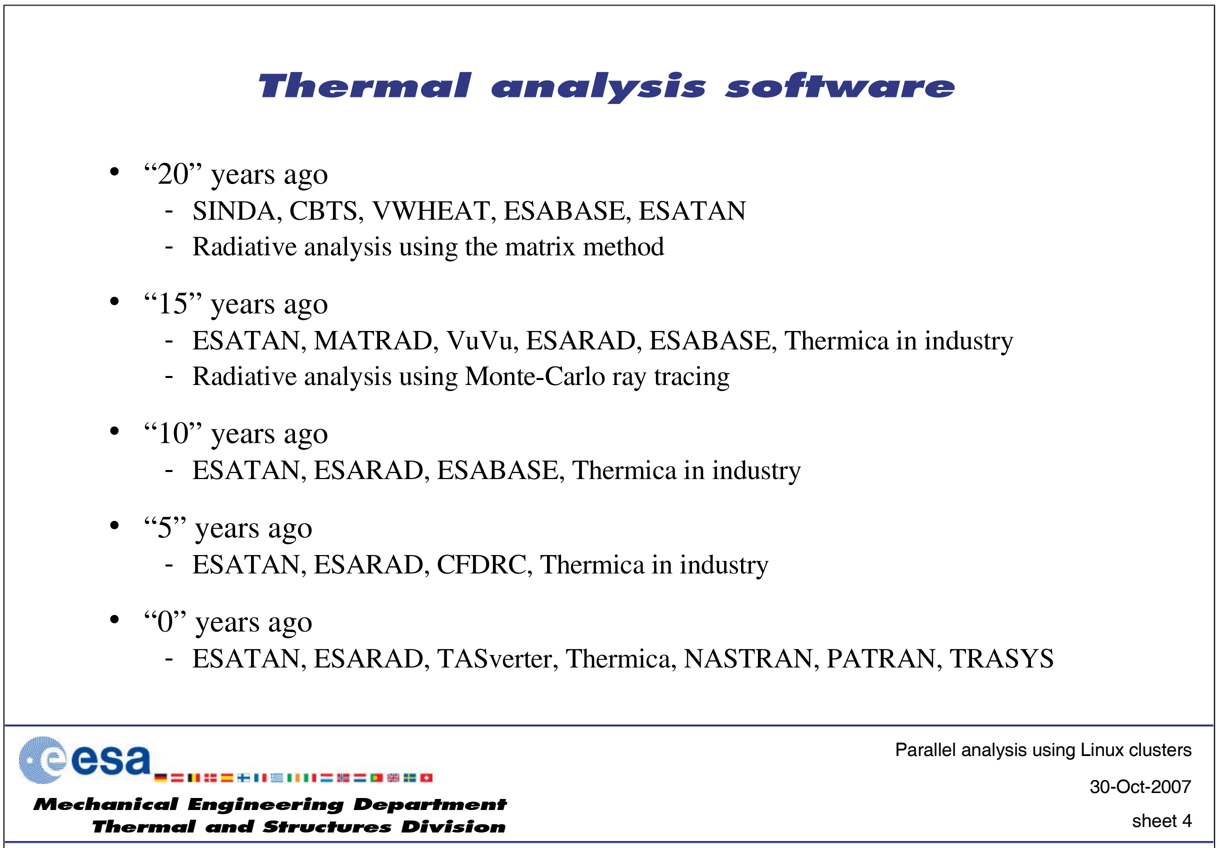
Move to a dedicated analysis facility in order to have some control and autonomy because thermal users had different (additional) requirements to office automation users . HP chosen because of positive experience of HP instrumentation and monitoring in the Mechanical System Laboratory. Only one graphics workstation for visualization. Later added Sun workstations, which were cheaper than HP, but also created compatibility problems.

Thin client X-terminals provide cheap access to graphical interfaces with central system administration compared to PCs or Unix workstations.

Mass purchase of cheaper PC hardware, with graphics, means users can use local processor for desktop work, leaving the central servers free for analysis. Initial Linux systems require installation and update on an individual basis.

Switch from HP to SGI servers due to harmonization with Structures group to streamline system administration. Switch from SGI to Sun after SGI’s future looked uncertain and their big servers were significantly more expensive than Sun. New systems are multi-processor.

“Unassigned” Linux desktops are put to use in an Linux cluster. Cluster grows and is renewed.



Thermal analysis software

- “20” years ago
  - SINDA, CBTS, VWHEAT, ESABASE, ESATAN
  - Radiative analysis using the matrix method
- “15” years ago
  - ESATAN, MATRAD, VuVu, ESARAD, ESABASE, Thermica in industry
  - Radiative analysis using Monte-Carlo ray tracing
- “10” years ago
  - ESATAN, ESARAD, ESABASE, Thermica in industry
- “5” years ago
  - ESATAN, ESARAD, CFDRC, Thermica in industry
- “0” years ago
  - ESATAN, ESARAD, TASverter, Thermica, NASTRAN, PATRAN, TRASYS

esa

Mechanical Engineering Department
Thermal and Structures Division

Parallel analysis using Linux clusters
30-Oct-2007
sheet 4

20 years ago, all tools – mostly text based - available on IBM mainframe or VAX/VMS.

Own environment more tailored for thermal analysis. Less rigid and more dynamic. More responsive to local users.

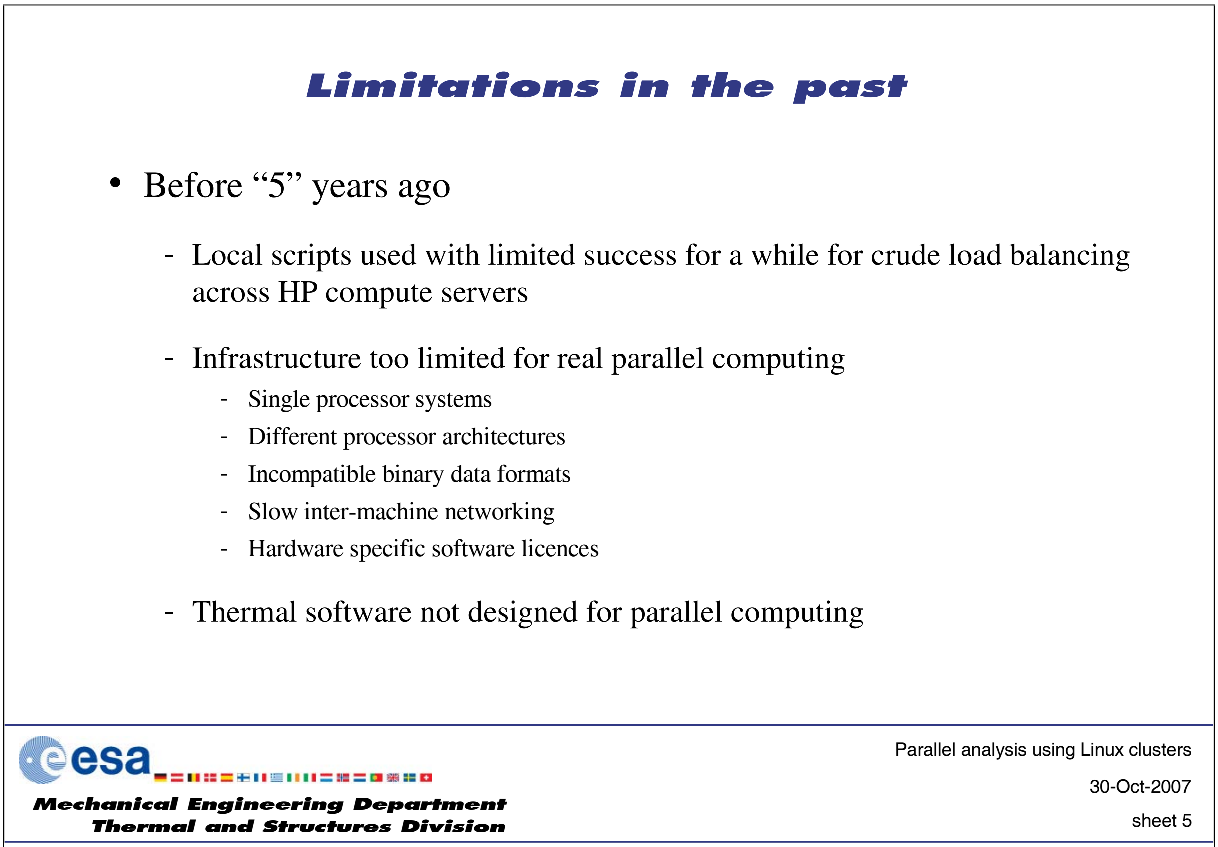
VuVu was an in-house tool for visualizing MATRAD geometry files. Tightly coupled to HP graphics accelerator hardware. Obsoleted by introduction of ESARAD on Sun workstations.

ESABASE used for “conversion” of Thermica models to ESARAD, but ran on VAX elsewhere in ESTEC.

Research Fellow joins the thermal division, who uses CFDRC to calculate convective air flows within ATV. Mesh adjusted so that various calculated values can be fed into ESATAN model.

In-house requirements for Thermica, NASTRAN, PATRAN and TRASYS as experience of data exchange for TASverter increases.

Interesting to note: 15 years ago 30 users shared 400Mb of disk space. Relatively simple ESARAD and ESATAN models. Typical ESARAD analysis runs all night. Now each user has 10Gb disk quota plus 50Gb of scratch space. Analysis requirements, ESARAD models and ESATAN models have become much more complex. Cryogenic models. High stability models. Typical ESARAD analysis still runs all night.



Limitations in the past

- Before “5” years ago
  - Local scripts used with limited success for a while for crude load balancing across HP compute servers
  - Infrastructure too limited for real parallel computing
    - Single processor systems
    - Different processor architectures
    - Incompatible binary data formats
    - Slow inter-machine networking
    - Hardware specific software licences
  - Thermal software not designed for parallel computing

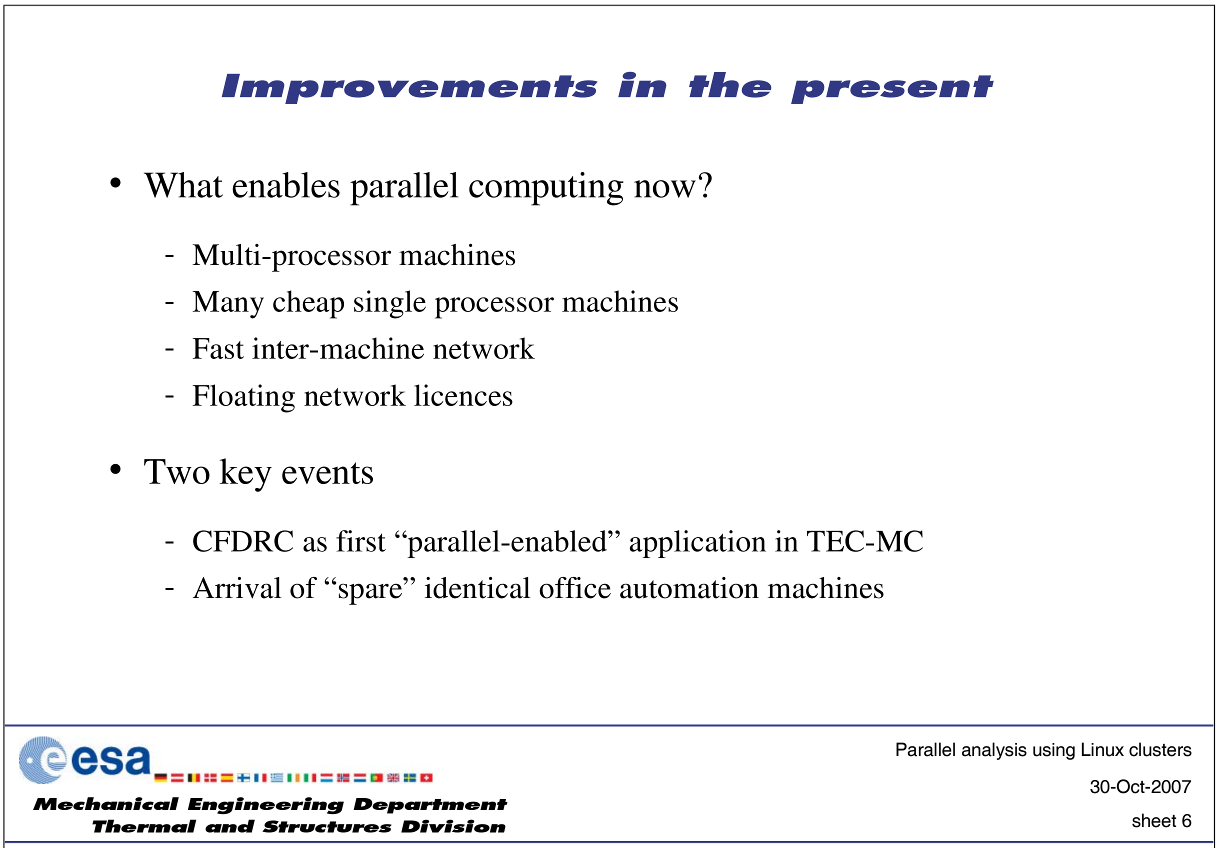
esa

Mechanical Engineering Department
Thermal and Structures Division

Parallel analysis using Linux clusters
30-Oct-2007
sheet 5

With 20 thermal engineers sharing the main compute server for their office automation work, each analysis job had a noticeable effect on system load and response times. Migrating one CPU intensive analysis job to the more lightly loaded graphics machine made a noticeable difference to performance.

But this requires having identical, or at least compatible systems, and/or using system independent data files, and having licences available on each system. Initial load balancing involved 2 HP800 series machines. Additional SUN systems could not be incorporated. Later 3 HP700 systems could run HP800 executables, but not the other way round. Much confusion among users about where jobs could and could not be run. Too much effort to maintain scripts to reflect differences and limitations on each machine.



Improvements in the present

- What enables parallel computing now?
  - Multi-processor machines
  - Many cheap single processor machines
  - Fast inter-machine network
  - Floating network licences
- Two key events
  - CFDRC as first “parallel-enabled” application in TEC-MC
  - Arrival of “spare” identical office automation machines

esa

Mechanical Engineering Department
Thermal and Structures Division

Parallel analysis using Linux clusters
30-Oct-2007
sheet 6

The rise in computer power has been matched by the increased complexity of the analysis but there are now other factors at play.

Multi-processor systems allow automatic load balancing by the system, so a single analysis job doesn’t overload the system and users can still work.

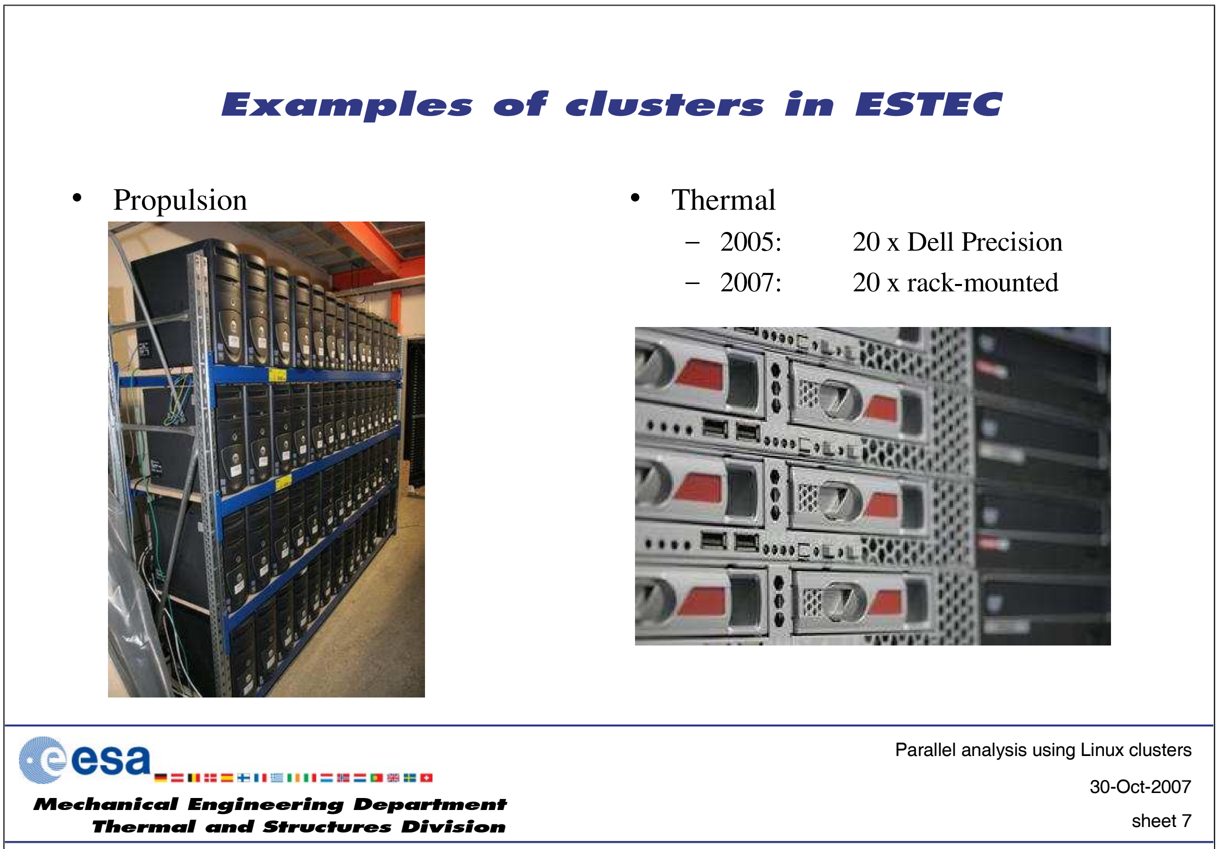
Mass produced standard PC hardware is much cheaper to buy than traditional minicomputers or workstations.

Faster networks allow much better data sharing between systems. Intelligent network switches provide better use of bandwidth.

Floating network licence managers mean that software can be made available across more systems without additional administration and configuration. Systems can be swapped in and out of the computing environment easily.

CFDRC was already geared up to work in a multi-processor environment. CFDRC software is computationally intensive and would be in constant use as each run could take days or even weeks. The ESARAD and ESATAN thermal analysis was not as intensive, with peaks of activity and periods of inactivity. The CFDRC user was keen to make maximum use of the system if the thermal analysis didn’t require it, but this proved difficult to organise. A turnkey system, with Linux and CFDRC on it was purchased. An experimental cluster was set up using “surplus” office automation PCs reconfigured for Linux.

Now CFDRC is being used in the Propulsions section on a dedicated 250-node Linux Cluster.



Examples of clusters in ESTEC

- Propulsion
- Thermal
  - 2005: 20 x Dell Precision
  - 2007: 20 x rack-mounted

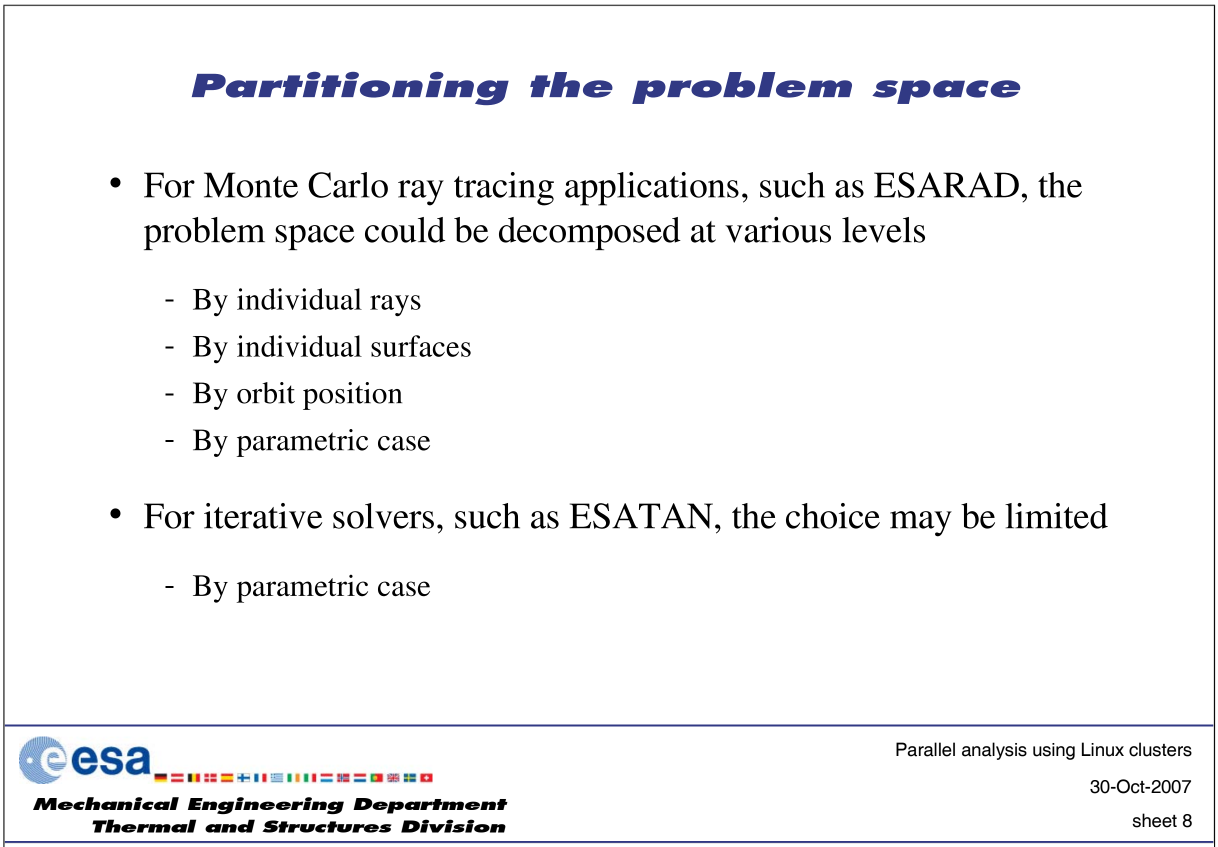
esa

Mechanical Engineering Department
Thermal and Structures Division

Parallel analysis using Linux clusters

30-Oct-2007

sheet 7



Partitioning the problem space

- For Monte Carlo ray tracing applications, such as ESARAD, the problem space could be decomposed at various levels
  - By individual rays
  - By individual surfaces
  - By orbit position
  - By parametric case
- For iterative solvers, such as ESATAN, the choice may be limited
  - By parametric case

esa

Mechanical Engineering Department
Thermal and Structures Division

Parallel analysis using Linux clusters

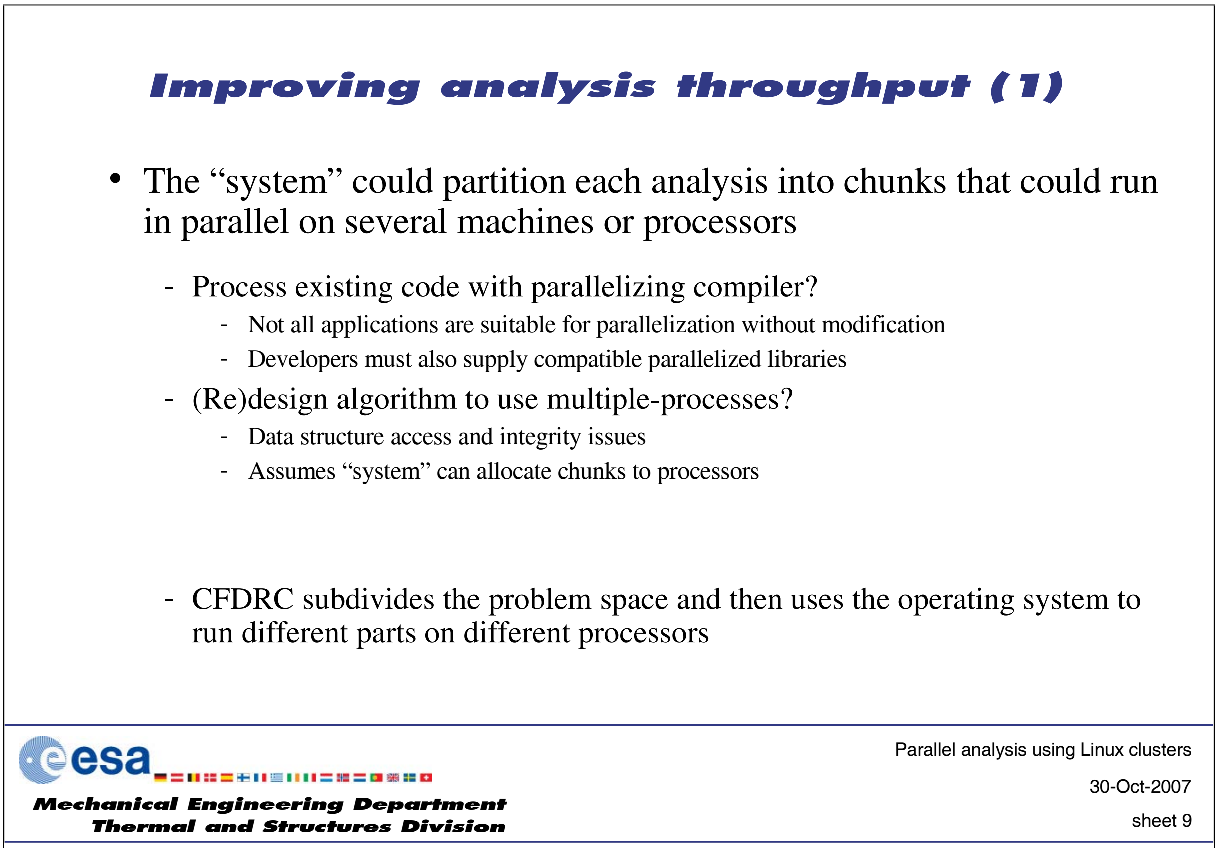
30-Oct-2007

sheet 8

It might be possible, at a conceptual level, to divide the problem space into different levels, but in reality to do so might require additional support from the application itself. If the level is too detailed, the overhead of running each smaller analysis might outweigh any benefits.

An ESATAN model could also be partitioned by splitting it into submodels or components that could be solved separately and having a management layer that handles communication of input/output values between the components. However, this presentation is about running models without major changes.

Note that tools that enable Stochastic analysis might provide a framework for running jobs in parallel across multiple processors but this is outside my experience. I haven't looked into it, but I am also curious whether the new ESATAN parametric case handling can be used in a parallel mode.



Improving analysis throughput (1)

- The “system” could partition each analysis into chunks that could run in parallel on several machines or processors
  - Process existing code with parallelizing compiler?
    - Not all applications are suitable for parallelization without modification
    - Developers must also supply compatible parallelized libraries
  - (Re)design algorithm to use multiple-processes?
    - Data structure access and integrity issues
    - Assumes “system” can allocate chunks to processors
  - CFDRC subdivides the problem space and then uses the operating system to run different parts on different processors

esa

Mechanical Engineering Department
Thermal and Structures Division

Parallel analysis using Linux clusters
30-Oct-2007
sheet 9

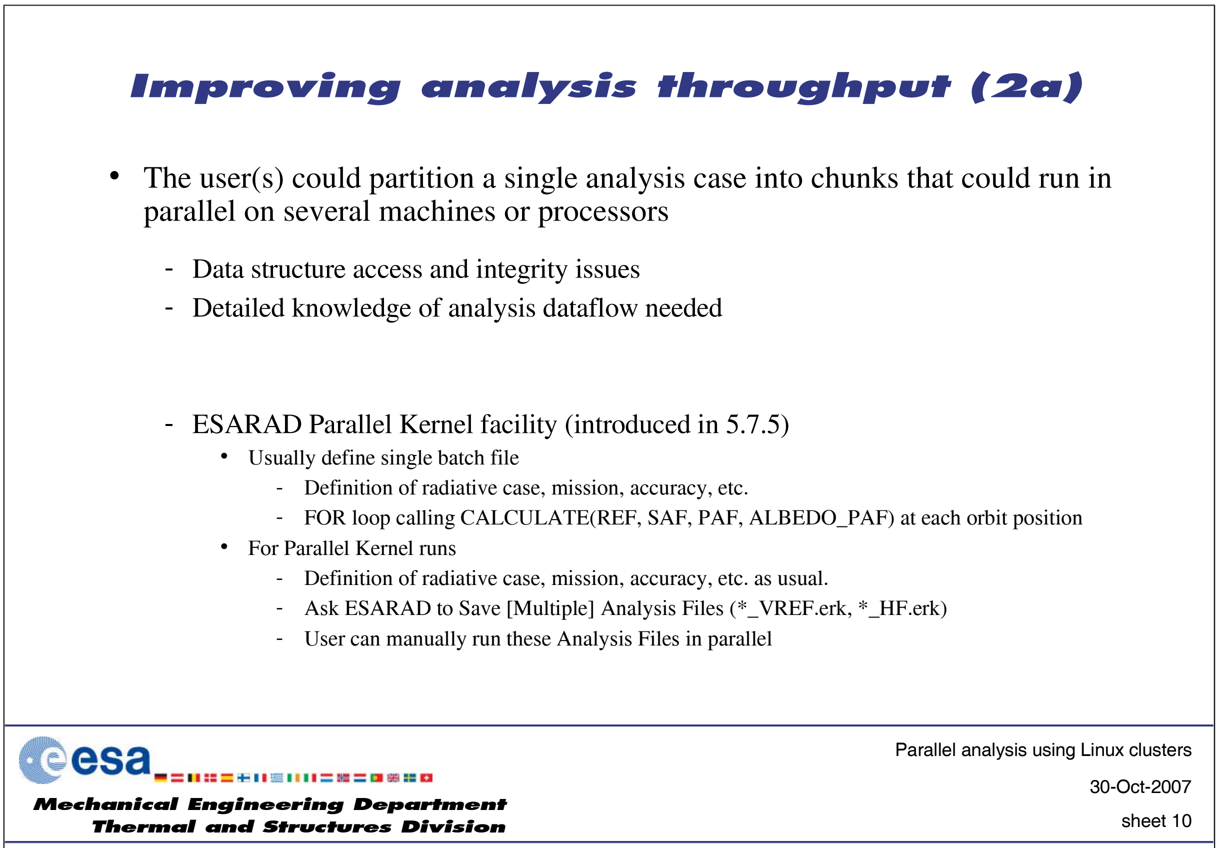
The ideal scenario for the user is for the “system” to do everything automatically: break the analysis into chunks, identify processors on which to run those chunks, and allocate chunks to processors. Unfortunately such a “system” does not really exist. Such a system can be broken down into three aspects: the multiple hardware processors, operating system support for job management on those processors, software support to work with the job management system.

Work may be required to rearrange algorithms and data flow in order to benefit from a parallelizing compiler.

Redesigning an algorithm to be multi-threaded can be a major task because there are critical issues with shared memory access and synchronization. Changing to use multiple processes involves additional effort in re-reading of input files and state information, inter-processor communication, and updating state information and integrating results. Care needs to be taken with data integrity in the event that a process is interrupted or fails.

A multi-processor system might provide automatic allocation of different chunks of analysis to the different local processors. Allocation of chunks across the network to remote processors requires some infrastructure support and configuration.

CFDRC provides most of the “system” for partitioning the complete analysis. The user needs to specify which processors are available for the computation. The CFDRC software allocates chunks of work to each processor. CFDRC is able to restart from a previous state if one of the processes fails or is interrupted.



Improving analysis throughput (2a)

- The user(s) could partition a single analysis case into chunks that could run in parallel on several machines or processors
  - Data structure access and integrity issues
  - Detailed knowledge of analysis dataflow needed

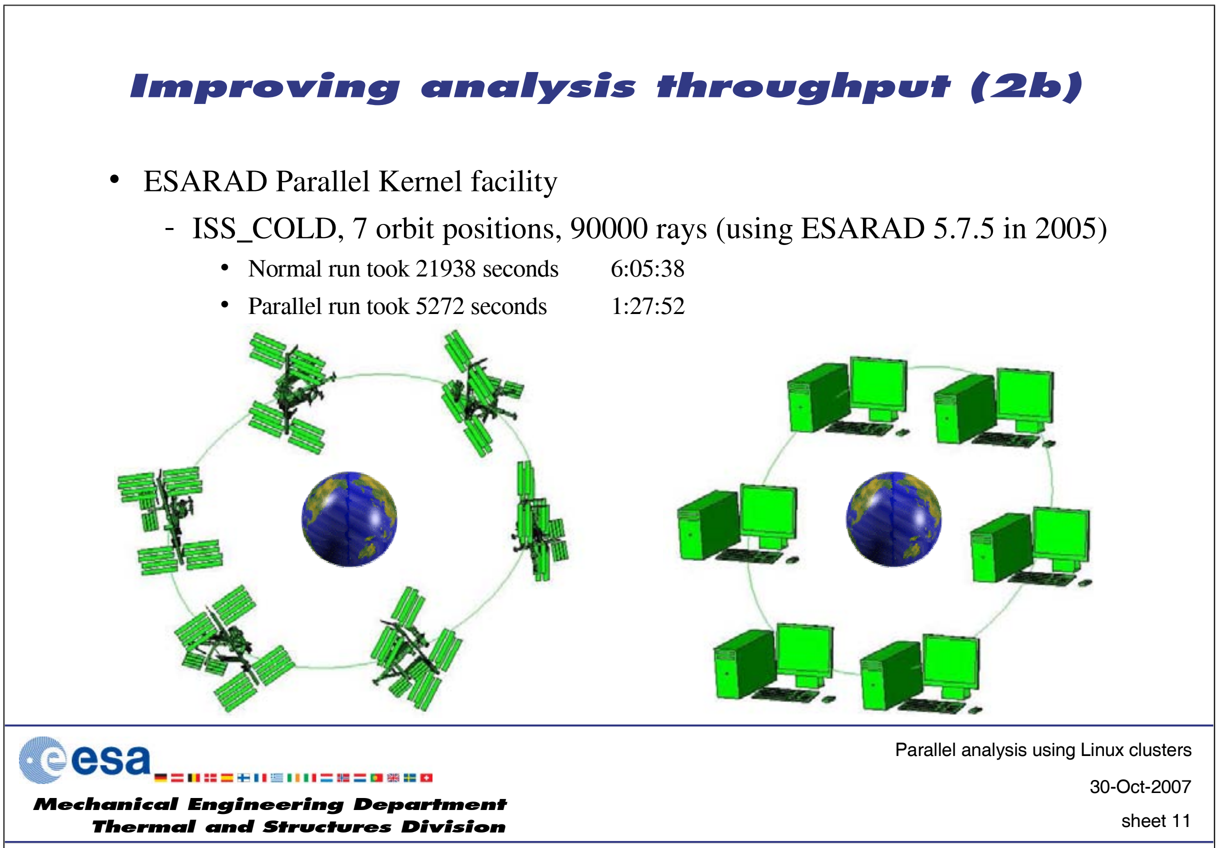
  - ESARAD Parallel Kernel facility (introduced in 5.7.5)
    - Usually define single batch file
      - Definition of radiative case, mission, accuracy, etc.
      - FOR loop calling CALCULATE(REF, SAF, PAF, ALBEDO_PAF) at each orbit position
    - For Parallel Kernel runs
      - Definition of radiative case, mission, accuracy, etc. as usual.
      - Ask ESARAD to Save [Multiple] Analysis Files (*_VREF.erk, *_HF.erk)
      - User can manually run these Analysis Files in parallel

esa

Mechanical Engineering Department
Thermal and Structures Division

Parallel analysis using Linux clusters
30-Oct-2007
sheet 10

Once the user has defined a satellite geometry and specified an orbit and other analysis parameters in a radiative case, Esarad's Parallel Kernel facility can be used to output individual kernel batch files that relate to calculating view factors or radiative exchange factors at one or more orbit positions and others for calculating the heat fluxes at those positions.



Improving analysis throughput (2b)

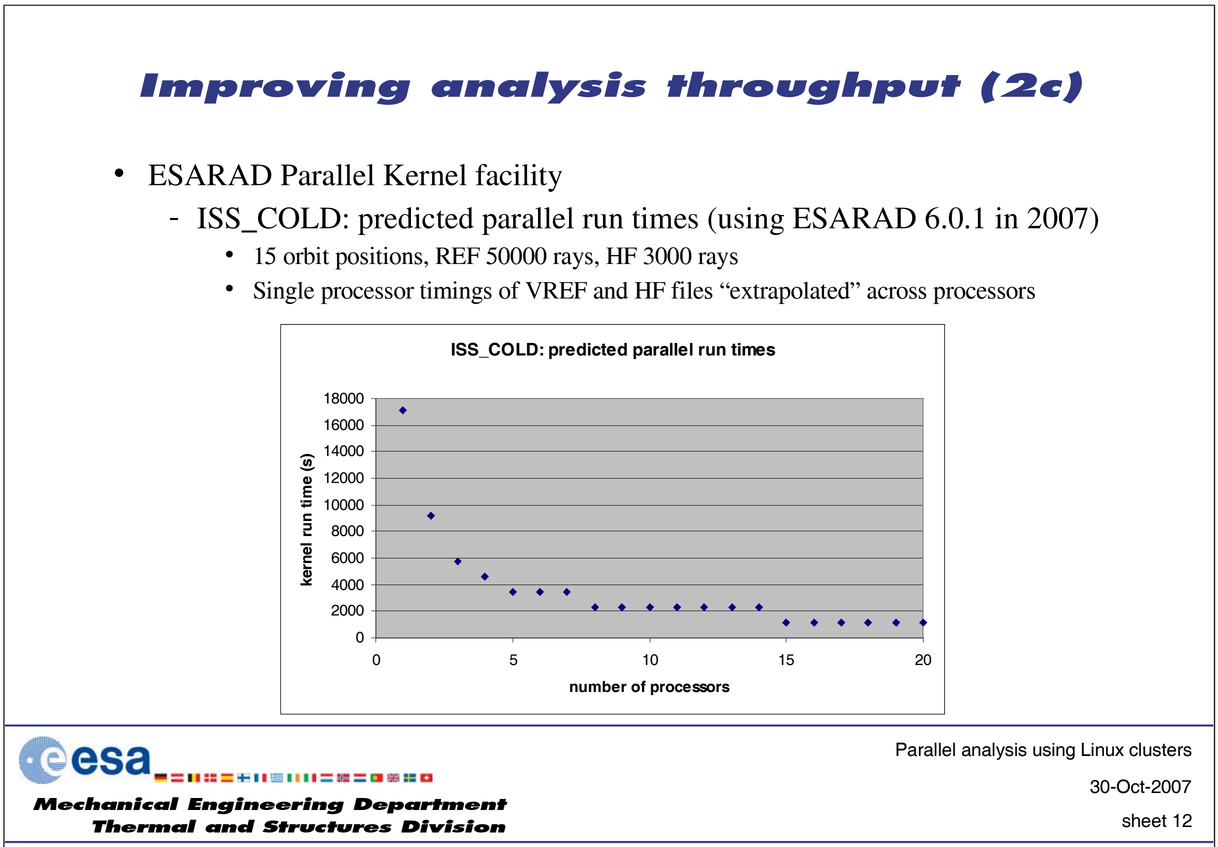
- ESARAD Parallel Kernel facility
  - ISS_COLD, 7 orbit positions, 90000 rays (using ESARAD 5.7.5 in 2005)
    - Normal run took 21938 seconds 6:05:38
    - Parallel run took 5272 seconds 1:27:52

esa

Mechanical Engineering Department
Thermal and Structures Division

Parallel analysis using Linux clusters
30-Oct-2007
sheet 11

In the example, the accumulated orientation of all of the surfaces in the ISS is different at each orbit position, so each position requires both a VF/REF calculation and a heat flux calculation. All machines share a networked file system and a shell script managed allocating and running each job using a remote shell command line and then waiting for the creation of a file showing that the batch job had completed. Synchronization was crude, with delays of 60 seconds between each job and while waiting for jobs to complete. Can see that there is some overhead because total time for parallel runs is greater than time for normal run divided by number of processors.



Improving analysis throughput (2c)

- ESARAD Parallel Kernel facility
  - ISS_COLD: predicted parallel run times (using ESARAD 6.0.1 in 2007)
    - 15 orbit positions, REF 50000 rays, HF 3000 rays
    - Single processor timings of VREF and HF files “extrapolated” across processors

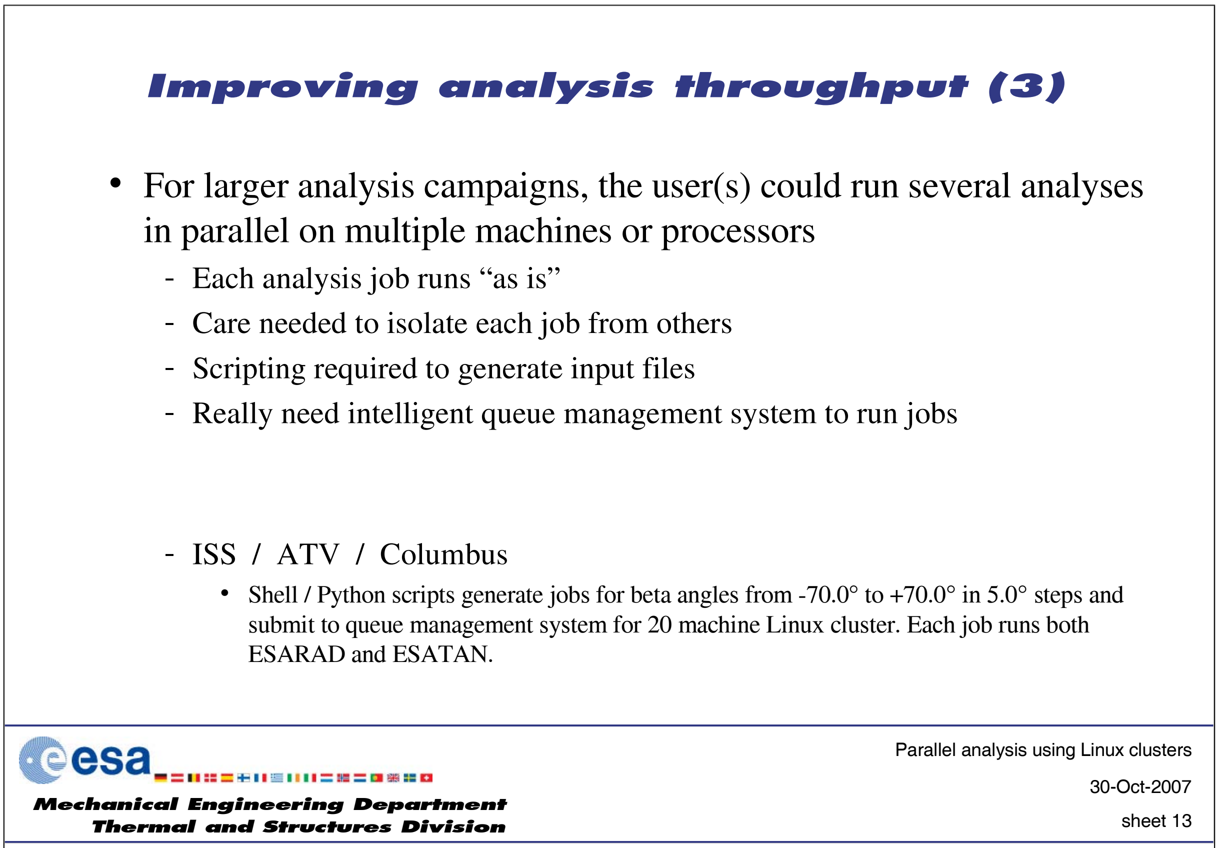
esa

Mechanical Engineering Department
Thermal and Structures Division

Parallel analysis using Linux clusters
30-Oct-2007
sheet 12

Unexpected interruptions in plans to upgrade hardware and software on the Linux cluster result in NFS problems that prevent using cluster machines to get up-to-date timing information for this presentation. The other option was a shared multi-processor server but this was already in heavy use for real project work using CFDRC. Therefore only able to predict parallel execution times based on results of running parallel jobs sequentially on one machine using a reduced number of rays. These are optimistic predictions because queue management and synchronisation overheads are not taken into account.

What is interesting is that it only takes a few processors to make a significant difference to the analysis time. So an engineer waiting for a job that would run overnight on a single processor could get results within the same working day if using 2, 3 or 4 processors. This is within the standard block of licences provided by Alstom, i.e. there is no need to buy additional licences to achieve a big improvement on turnaround time for a single analysis case.



Improving analysis throughput (3)

- For larger analysis campaigns, the user(s) could run several analyses in parallel on multiple machines or processors
  - Each analysis job runs “as is”
  - Care needed to isolate each job from others
  - Scripting required to generate input files
  - Really need intelligent queue management system to run jobs
  - ISS / ATV / Columbus
    - Shell / Python scripts generate jobs for beta angles from -70.0° to +70.0° in 5.0° steps and submit to queue management system for 20 machine Linux cluster. Each job runs both ESARAD and ESATAN.

esa

Mechanical Engineering Department
Thermal and Structures Division

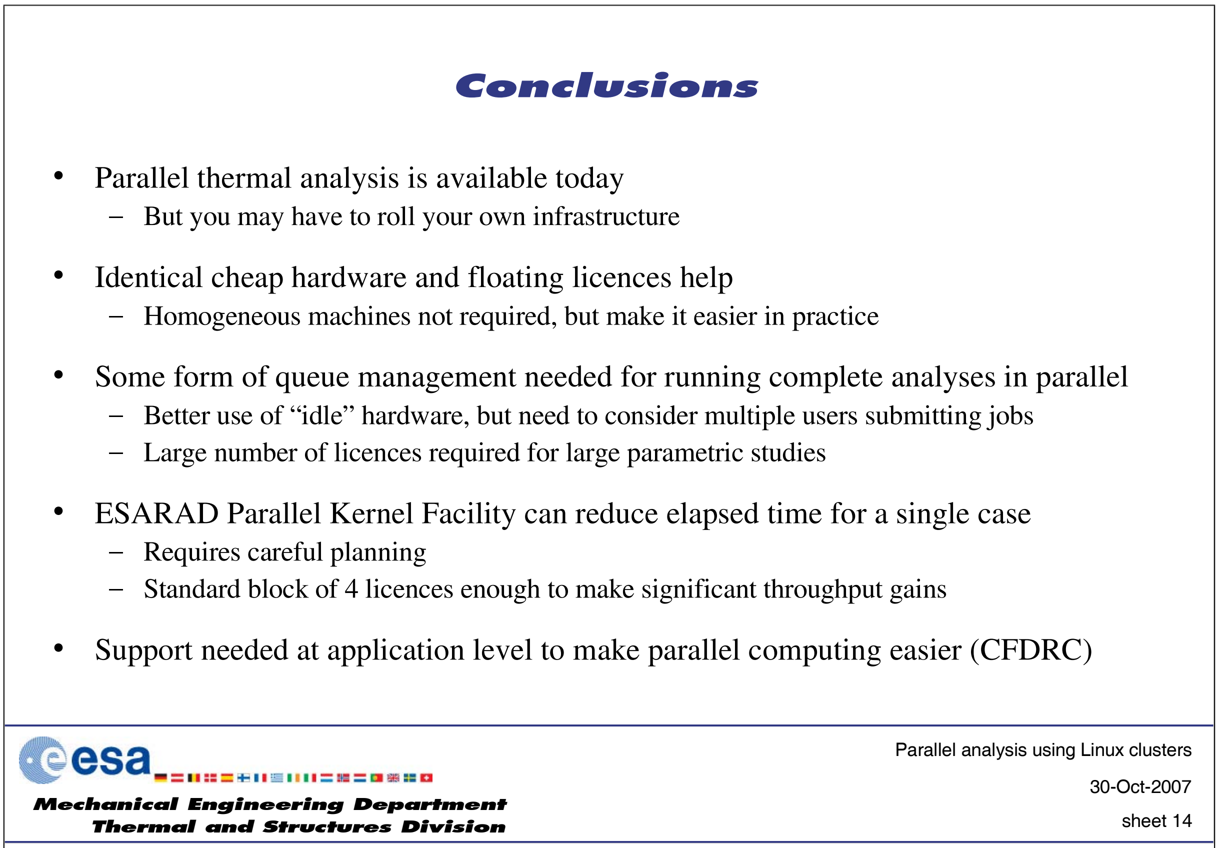
Parallel analysis using Linux clusters
30-Oct-2007
sheet 13

A set of template files created from original batch files.

A shell or python script used to read parametric data from file, and copy the template files to a new directory, substituting the parametric data in the process, and then generating another script to set the appropriate environment (ESARAD_HOME, etc) and then submit this second script to the queue management system. The advantage of the queue management system is that it handles all allocation and synchronization. The disadvantage is that you need to manage disk space so your second script needs to tidy up after itself. Queue management system needs to be configured so that the first user with 100s of jobs doesn’t block anyone else.

Management need to decide policy before it can be configured into such a system, i.e. number of simultaneous jobs per user, who has priority, etc.

Queue management system used was commercial tool called PBSPro. PBS stands for Portable Batch System. The original version was developed for NASA in the mid 1990s. An open source version called OpenPBS is also available, but is no longer under development. Details of both tools can be found via http://www.pbspro.com/openpbs.html



Conclusions

- Parallel thermal analysis is available today
  - But you may have to roll your own infrastructure
- Identical cheap hardware and floating licences help
  - Homogeneous machines not required, but make it easier in practice
- Some form of queue management needed for running complete analyses in parallel
  - Better use of “idle” hardware, but need to consider multiple users submitting jobs
  - Large number of licences required for large parametric studies
- ESARAD Parallel Kernel Facility can reduce elapsed time for a single case
  - Requires careful planning
  - Standard block of 4 licences enough to make significant throughput gains
- Support needed at application level to make parallel computing easier (CFDRC)

Mechanical Engineering Department
Thermal and Structures Division

Parallel analysis using Linux clusters
30-Oct-2007
sheet 14