ASCI Blue Mountain

Lars Kr. Lundin

- **CERFACS Student `96-`97**
- **LANL Post doc `99-2001**
- **ESO Staff member since 2003**

Presentation Outline

- ASCI + LANL/Algo-team
- ASCI Blue Mountain description
- ASCI Blue Mountain first evaluation
- ASCI Blue Mountain second evaluation
- ESO
- ESO VLT data processing

ASCI + LANL/Algo-team

Accelerated Strategic Computing Initiative after 1992 US nuclear testing moratorium.

Aim of adding 5 orders of magnitude to simulation capability of stockpile stewardship.

Partly by building new, powerful one-of-a-kind supercomputers and
partly by finding algorithms that would work well on these computers.

First two ASCI computers ASCI Red + Blue Mountain operational in `97 + `98.
- the first to break the 1 TFLOP/s limit on the LINPACK benchmark.

Parallel Architectures and Algorithms Team provided algebraic multigrid
and Krylov solvers to 3D-modelling projects at LANL (radiation transport).

The essential component a parallel, distributed sparse matrix-vector product.

ASCI Blue Mountain Description

Actually two separate clusters of Origin 2000 SMPs (“boxes”):

- One open cluster of 16 HIPPI connected boxes.
- One restricted cluster of 48 HIPPI connected boxes

Each box comprises 128 MIPS R10000 250MHz CPUs:

- Each CPU has a theoretical peak-performance of 500MFLOP/s.
- CPUs mounted in pairs on one board, 512MB RAM per board.
- The 64 boards connected in pairs as vertices in a 5-dimensional hypercube
- 32GB RAM and up to 64GFLOP/s per box.

In total:

- Open: 2048 CPUs w. 512GB RAM, up to 1024GFLOP/s.
- Restricted: 6144 CPUs w. 1536GB RAM, up to 3072GFLOP/s.

ASCI Blue Mountain Usage

Open cluster batch system:

- Daytime *scavenger* queue (up to 1024 CPUs for up to 15 minutes).
- Largest queue: 2048 CPUs for up to 8 hours (One night per week only).

Means that:

- 1024 CPU runs were routine (on daily basis).
- 2048 CPU runs done once every few weeks.
- 6144 CPU runs possible only via delivery to X-division

Billing policy:

- Simply multiply wall-clock time with number of processors
  - although fair, does not promote parallelism.
- I was first in dept. to make an all-night run on 2048 CPUs,
  - equivalent to 3 months of projected CPU usage...

ASCI Blue Mountain First Evaluation

Find characteristics of this leading-edge one-of-a-kind computer.

First evaluation using familiar 3D CFD code,

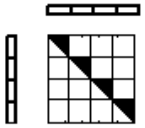
- Conjugate Gradient (Least Squares) on a sparse, banded matrix.
- Parallel CGLS Solver written at CERFACS
- MPI v. 1.1 with a pipe topology
- Portable FORTRAN-77

PCGLS
- on four processors

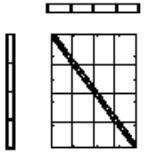
IMM

PCGLS:

a) $t^{(k)} - S^{-1}p^{(k)}$



b) $q^{(k)} - At^{(k)}$



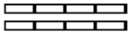
Send to left neighbor

c) $\alpha_k - \frac{\gamma_k}{q^{(k)T}q^{(k)}}$

Global sum

d) $x^{(k+1)} - x^{(k)} + \alpha_k t^{(k)}$

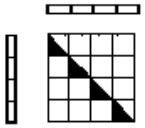
e) $r^{(k+1)} - r^{(k)} - \alpha_k q^{(k)}$



f) $\tilde{s}^{(k)} - A^T r^{(k-1)}$

Send to right neighbor

g) $s^{(k-1)} - S^T \tilde{s}^{(k)}$



h) $\gamma_{k-1} - s^{(k+1)T} s^{(k+1)}$

Global sum

i) $\beta_k - \frac{\gamma_{k-1}}{\gamma_k}$

j) $p^{(k+1)} - s^{(k+1)} + \beta_k p^{(k)}$

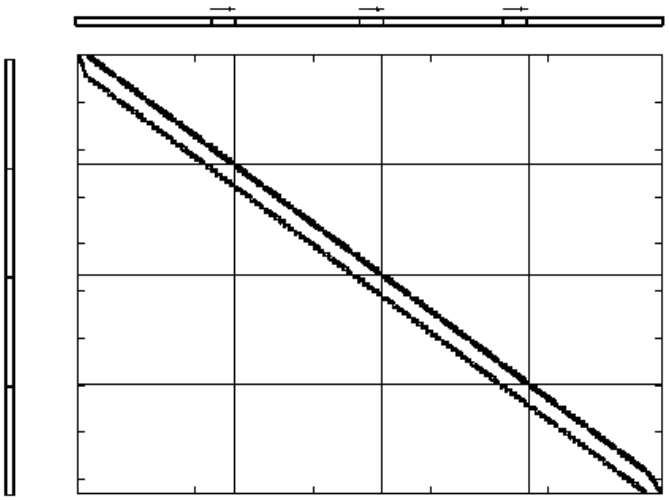
CERFACS January '97 lkl@imm.dtu.dk

Communication needed for matrix-vector products and for dot-products

Sparse, banded matrix vector product - on four processors

IMM

f) $\tilde{s}^{(k)} = A^T p^{(k-1)}$



Send to right neighbor

Receive from left neighbor

CERFACS January '97 lkl@imm.dtu.dk

Minimize communication overhead:

- **Non blocking communication**
  - try to overlap communication with computation.
- **Persistent messages**
  - reuse setup of messages sent in every iteration

CGLS solver Performance
- Relative solution time on 1.5E6 unknowns

½ Blue mountain box →

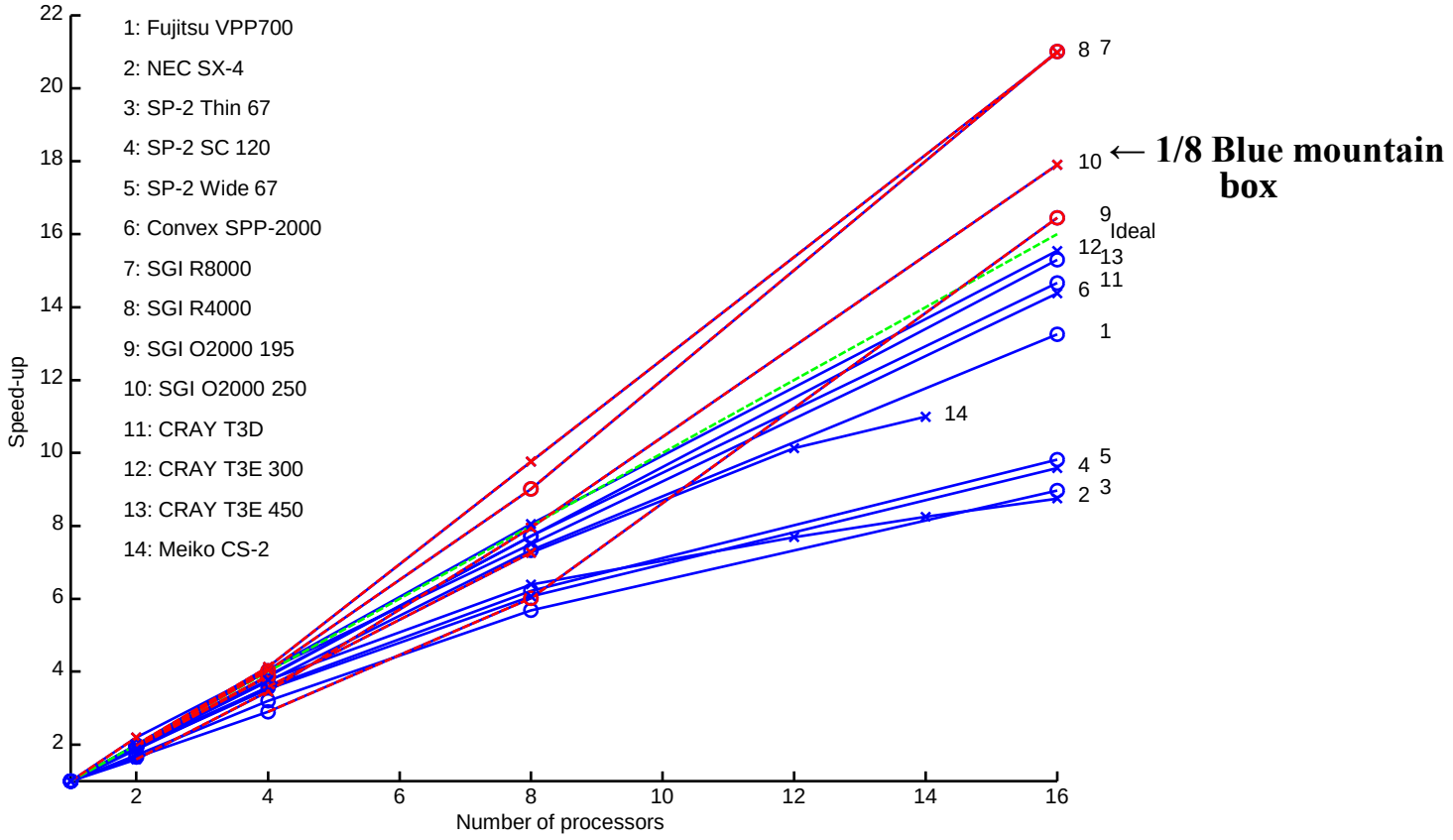
5000€ 2007 Server →

Computer	Short Name	CPU's	Peak [MFLOP/s]	T_n
Fujitsu VPP700	VPP	32	70400	1.00*
Cray T3E/450	T3E450	64	57600	2.75*
Cray T3E/300	T3E300	64	38400	3.67
NEC SX-4	NEC	16	35200	1.97*
SGI Origin 2000/250	O2K250	64	32000	1.41*
Cray T3D/150	T3D	64	19200	6.73
Intel Dual Xeon-4/2330	Xeon8	8	18640	4.51
IBM SP-2 SC/120	SP120	32	15360	4.08
SGI Origin 2000/195	O2K195	32	12480	2.68
Convex SPP-2000	SPP	16	11520	7.78
IBM SP-2 Thin/67	SPthin	32	8512	8.62*
SGI PC R8000	R8K	16	5760	5.84
IBM SP-2 Wide/67	SPwide	16	4256	8.57
SGI PC R4000	R4K	16	4000	11.26
Meiko CS-2	CS2	14	1400	22.67
SGI PC R10000	R10K	6	2328	23.90
CRAY T90	T90	1	1920	65.54
SUN HPC450	SUN	2	1000	45.67
CRAY C92A	C90	1	960	29.02
HP 9000/180	HP	1	720	166.24
IBM SP-2 SC/135	SP135	1	540	49.10
CRAY Y-MP	YMP	1	330	62.72
Pentium PRO/200	PC	1	200	370.76

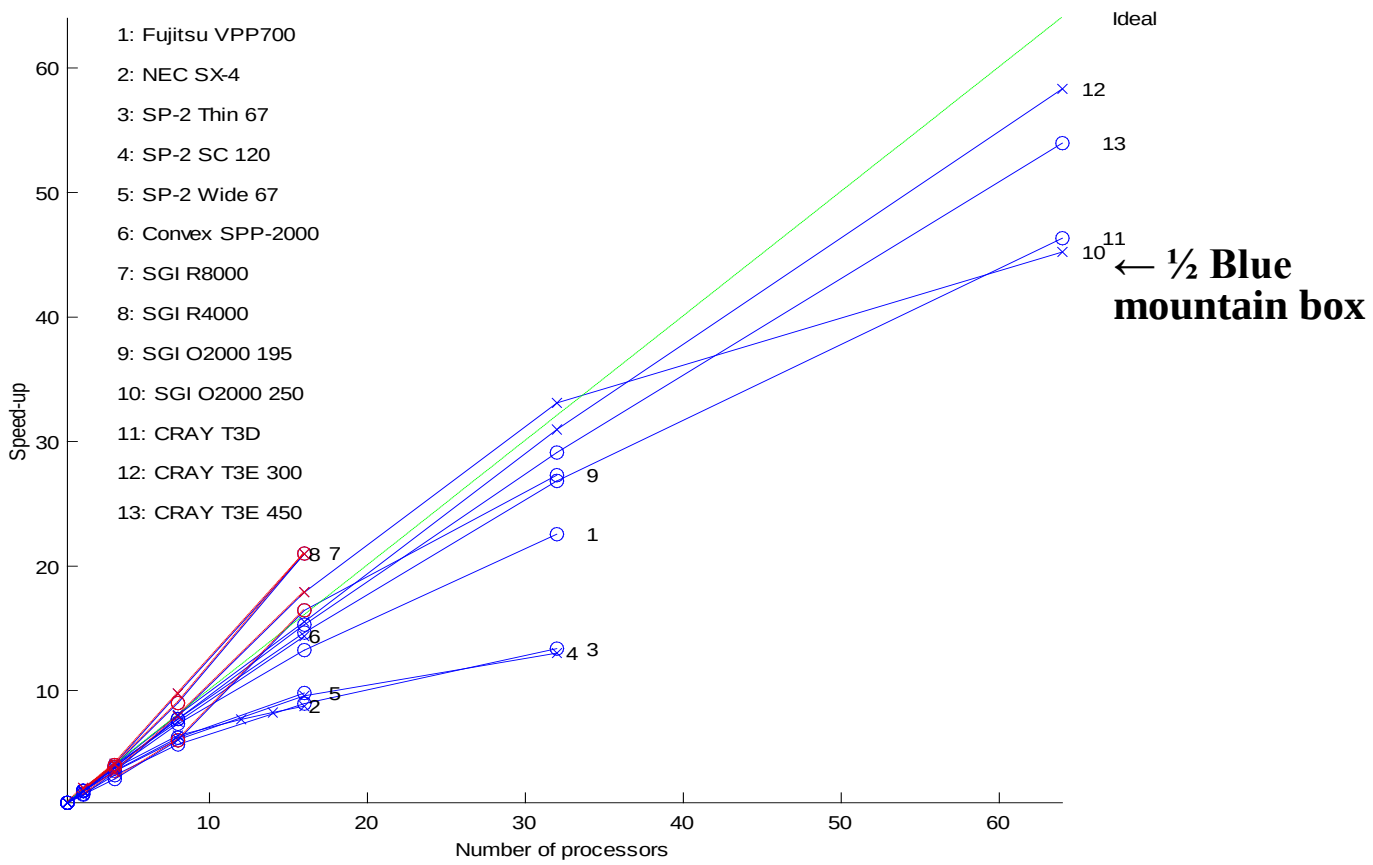
CGLS solver Performance
- Relative solution time on 1.5E6 unknowns

Computer	Short Name	CPU's	Peak [MFLOP/s]	T_n
Fujitsu VPP700	VPP	32	70400	1.00*
Cray T3E/450	T3E450	64	57600	2.75*
Cray T3E/300	T3E300	64	38400	3.67
NEC SX-4	NEC	16	35200	1.97*
SGI Origin 2000/250	O2K250	64	32000	1.41*
Cray T3D/150	T3D	64	19200	6.73
Intel Dual Xeon-4/2330	Xeon8	8	18640	4.51

CGLS solver Performance
- Parallel speed-up on 1.5E6 unknowns



CGLS solver Performance
- Parallel speed-up on 1.5E6 unknowns

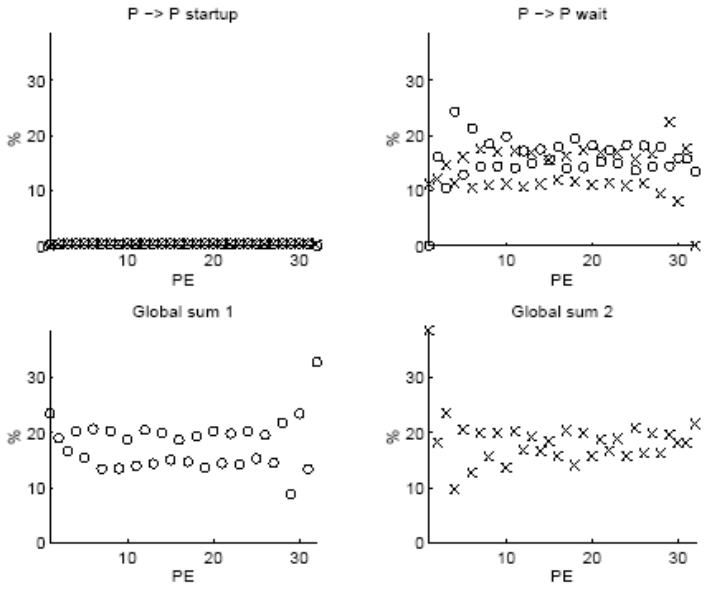


← ½ Blue mountain box

IMM

IBM SP-2 Thin

32 processors



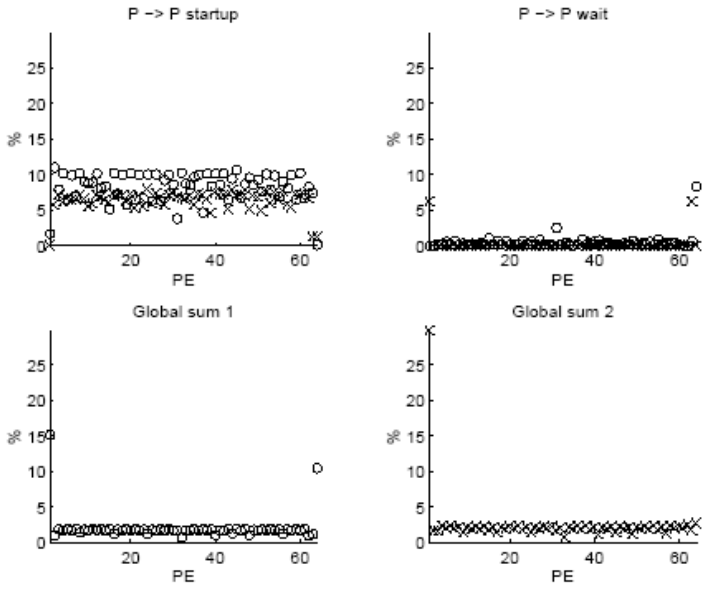
CERFACS January '97 lkl@imm.dtu.dk

Example:
No start-up cost compared to wait cost

IMM

Cray T3D

64 processors



Example:
No wait cost compared to start-up cost

CERFACS January '97 lkl@imm.dtu.dk

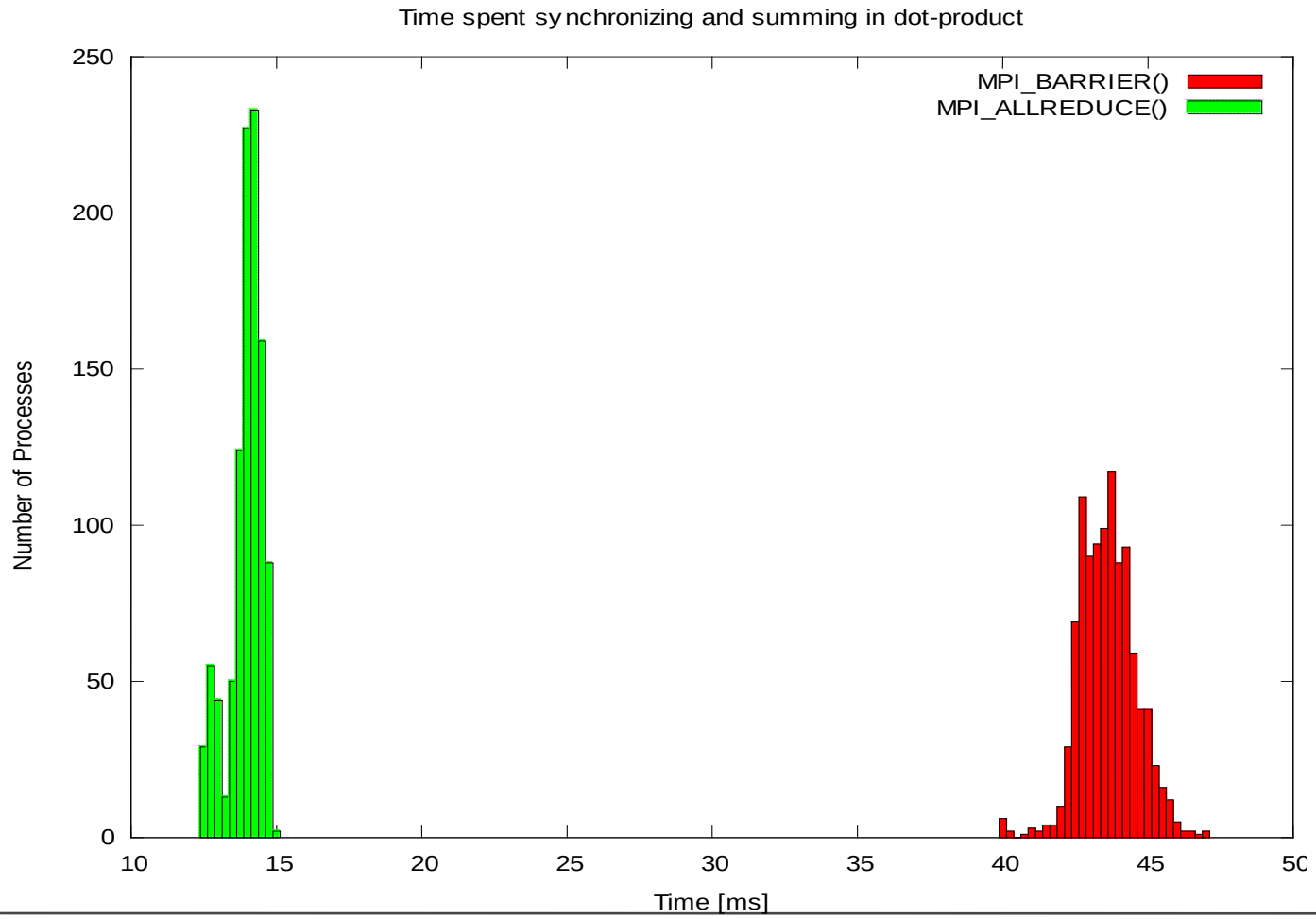
ASCI Blue Mountain Second Evaluation

The CGLS evaluation has too small problem size.
The discretization did not foresee thousands of processors.

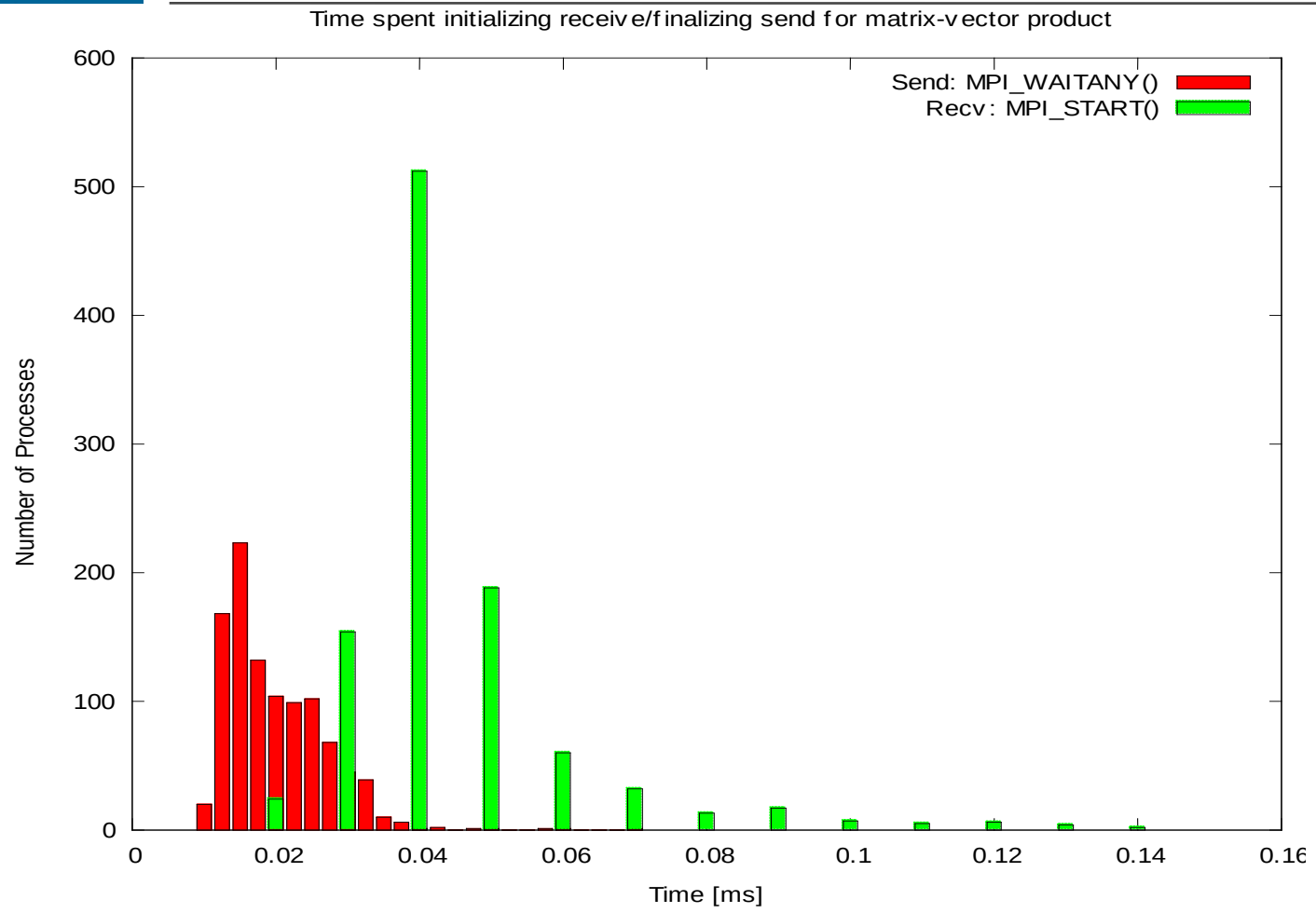
Write new solver for 3D radiation transport code:

- Conjugate Gradient on a sparse, banded SPD matrix.
- MPI v. 1.1 with a pipe topology
- Matrix-vector-product has non-local parts above and below main diagonal, thus bi-directional communication
- Map logical topology onto physical.
- Wait for first arriving message (from left/right)

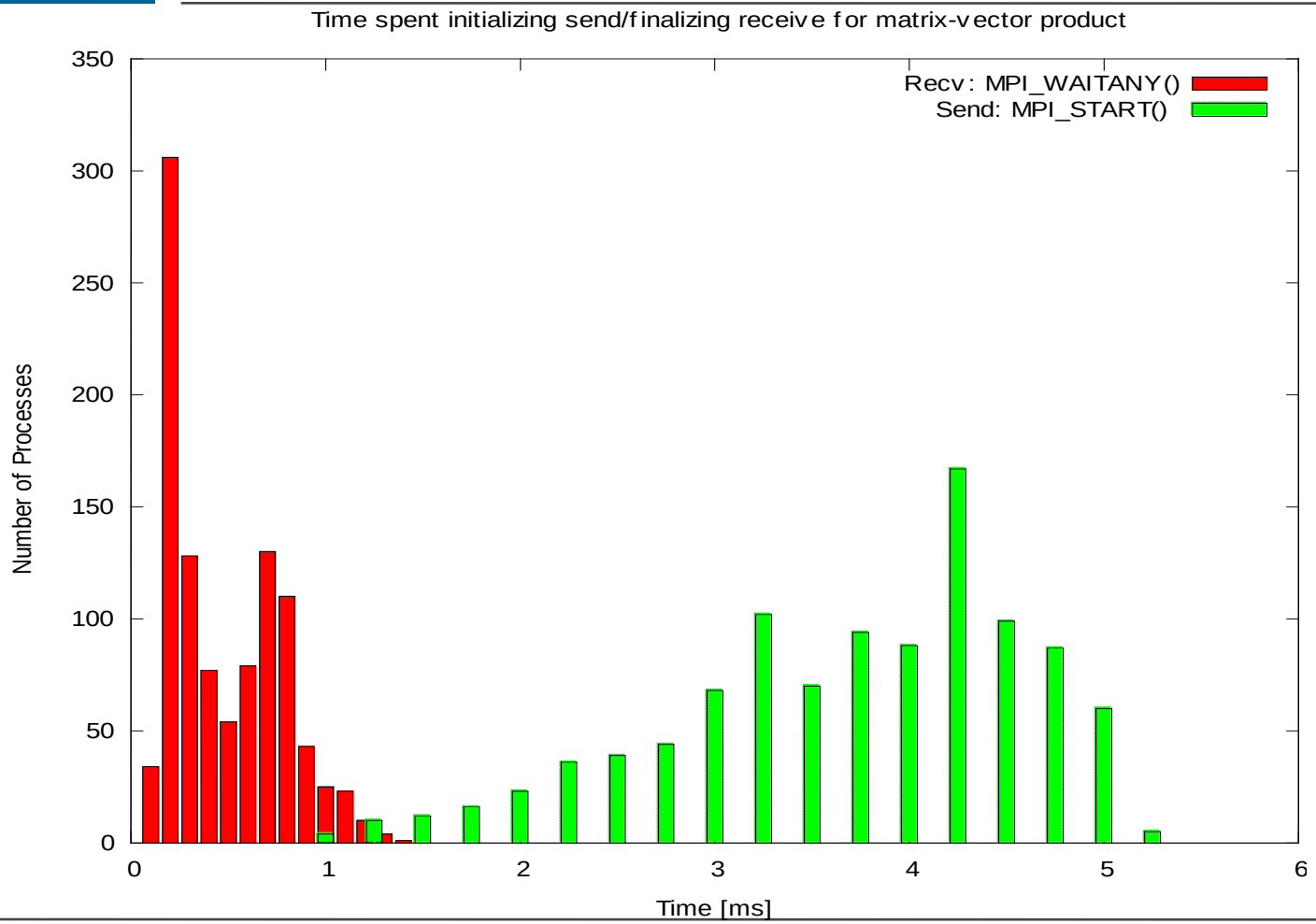
ASCI Blue Mountain CG evaluation 64E6 unknowns – 1024 Processors



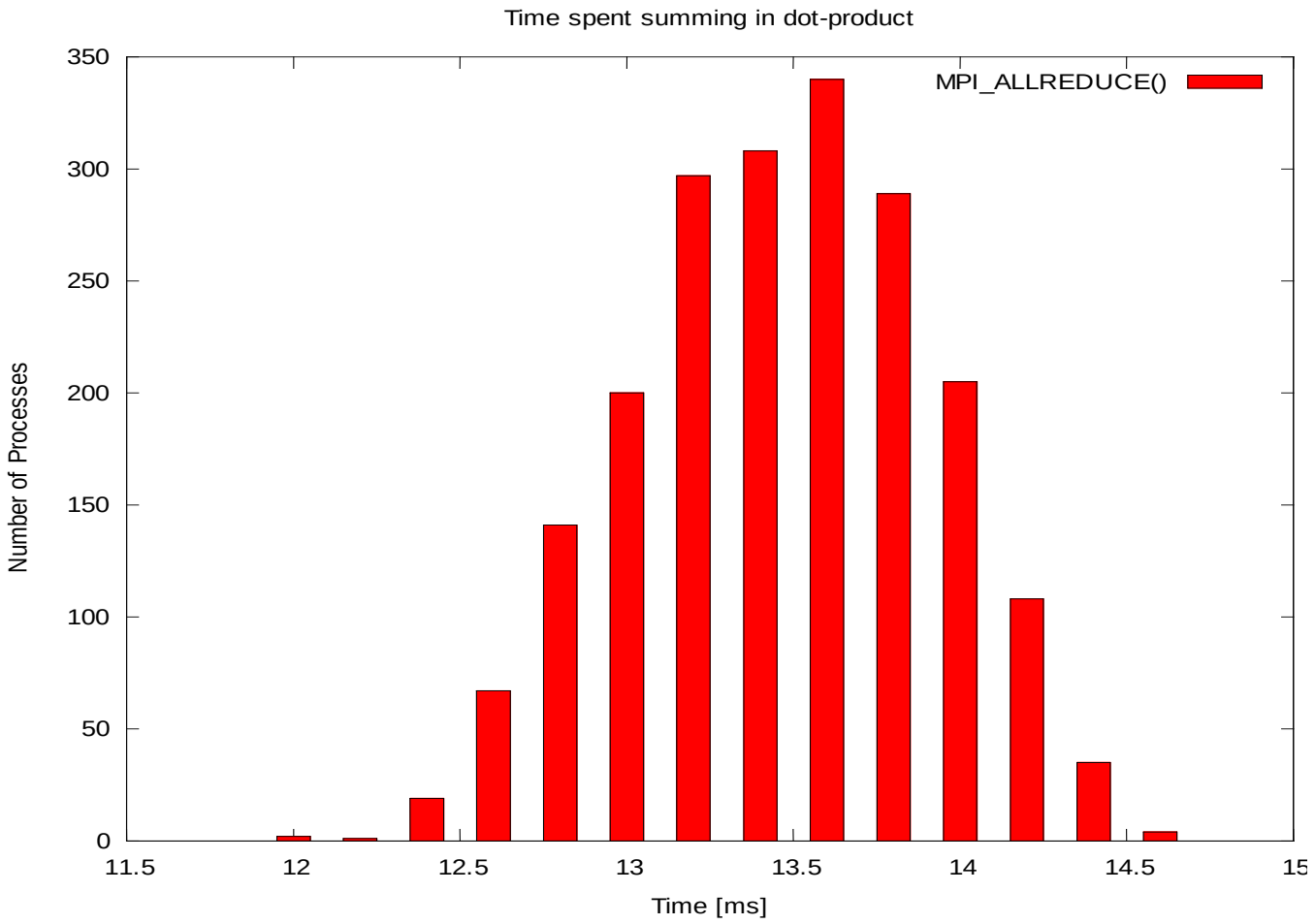
ASCI Blue Mountain CG evaluation 64E6 unknowns – 1024 Processors



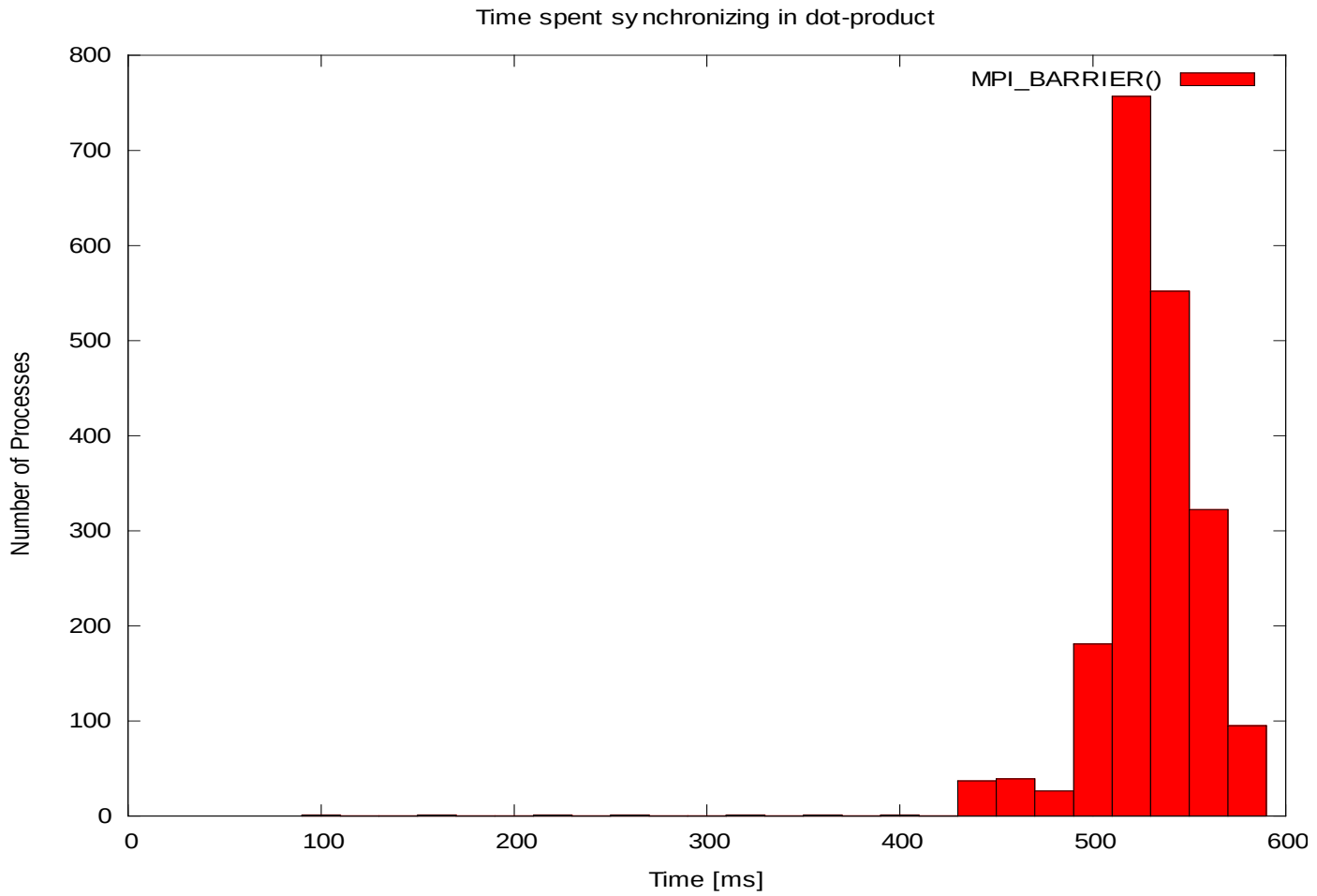
ASCI Blue Mountain CG evaluation 64E6 unknowns – 1024 Processors



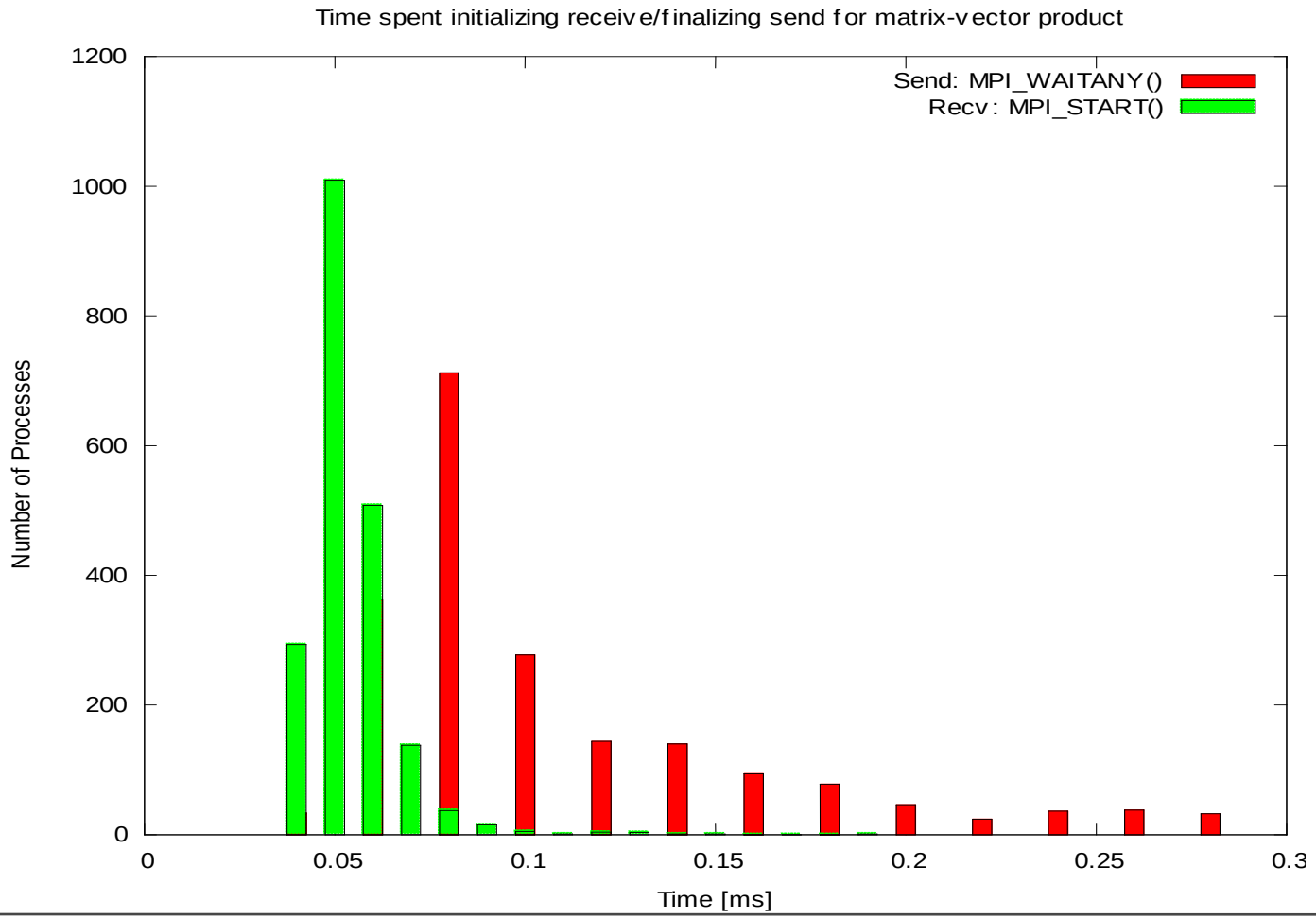
ASCI Blue Mountain CG evaluation
1E9 unknowns – 2048 Processors



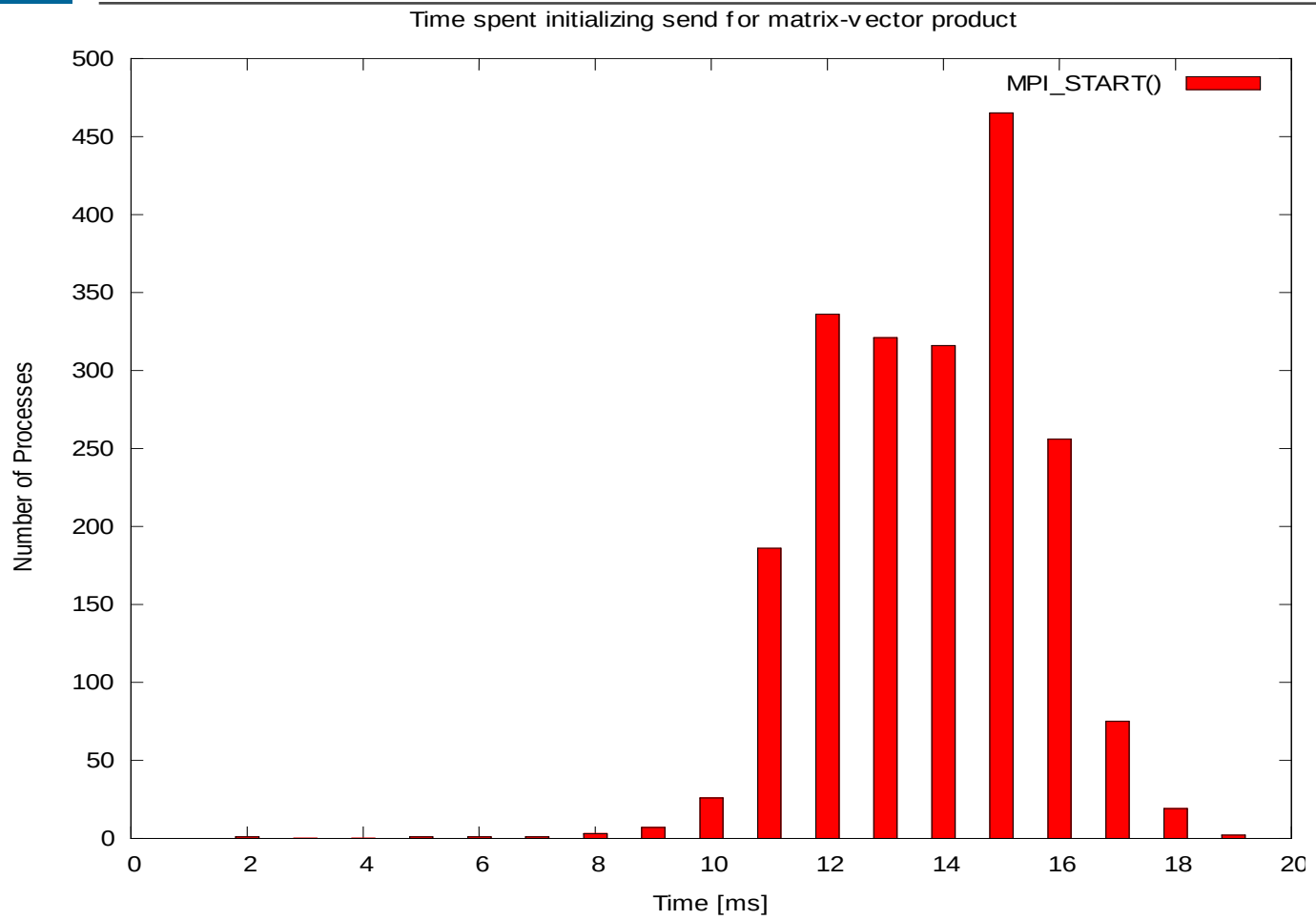
ASCI Blue Mountain CG evaluation
1E9 unknowns – 2048 Processors



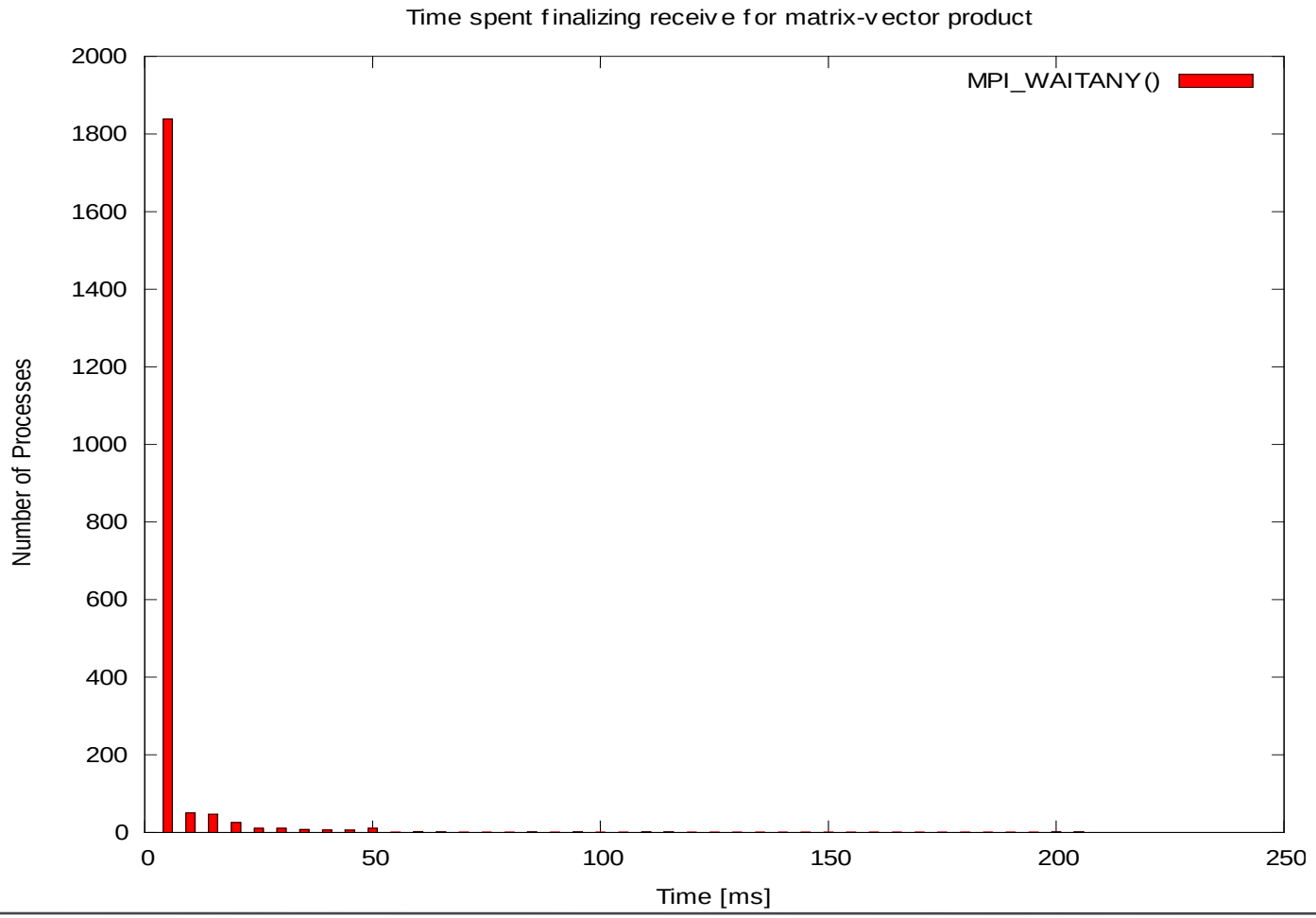
ASCI Blue Mountain CG evaluation
1E9 unknowns – 2048 Processors



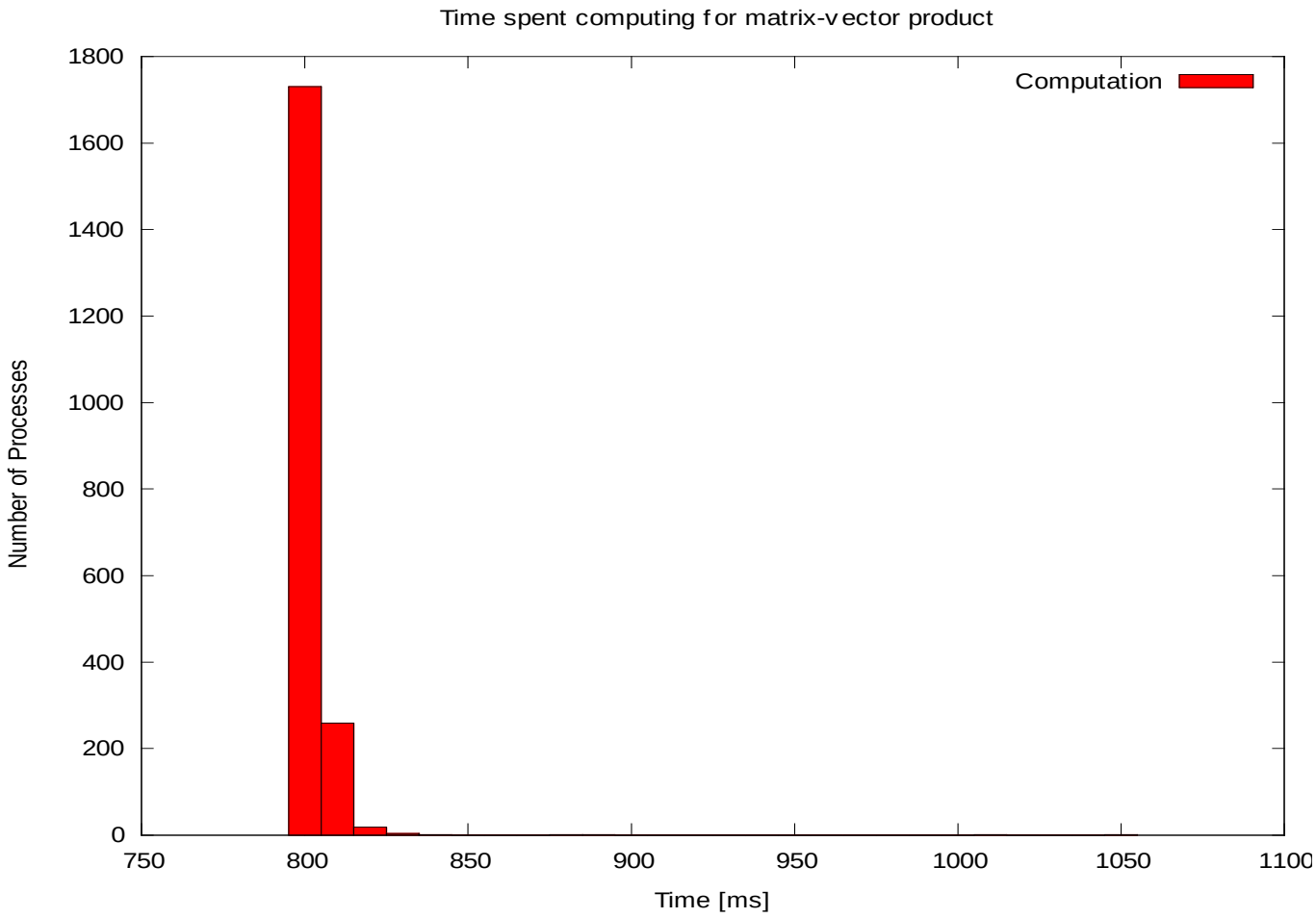
ASCI Blue Mountain CG evaluation
1E9 unknowns – 2048 Processors



ASCI Blue Mountain CG evaluation
1E9 unknowns – 2048 Processors



ASCI Blue Mountain CG evaluation
1E9 unknowns – 2048 Processors



Blue Mountain Conclusions

Blue Mountain Performance:

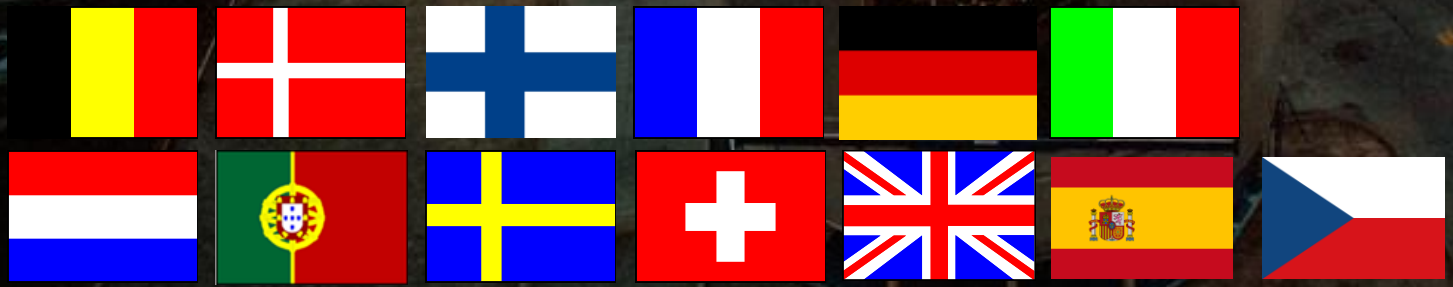
- **2048 processor efficiency at 65% (3D problem)**
- **2048 processor efficiency at 75% (2D problem)**
- **1024 processor efficiency at 90% (“small” 3D problem)**
- **Limited efficiency due to system load on two CPUs**
- **Any gain from global sum optimizations likely lost.**

Main lesson learned:

- **Measurement better than assumption, especially with ‘bleeding edge’ technology.**

ESO is…

- **Intergovernmental Organisation**
- **International Convention (1962)**



ESO's mission…

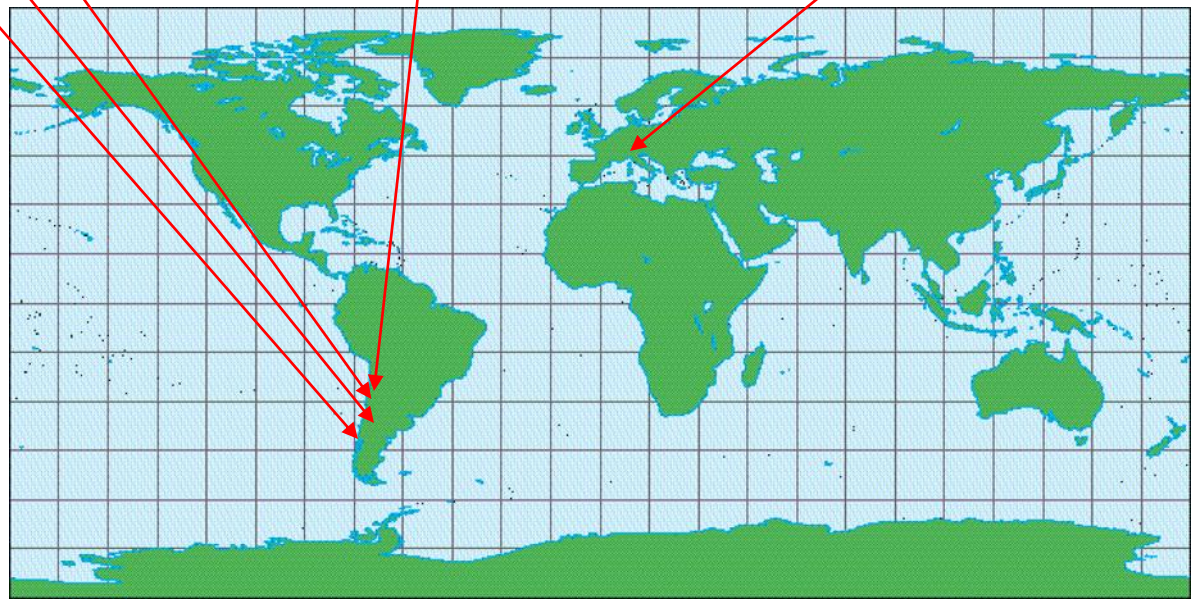
- **Provide astronomers in member states (~ 3000) with state-of-the-art observational facilities**
- **Further and organise collaboration with astronomy and astro-physics in Europe**

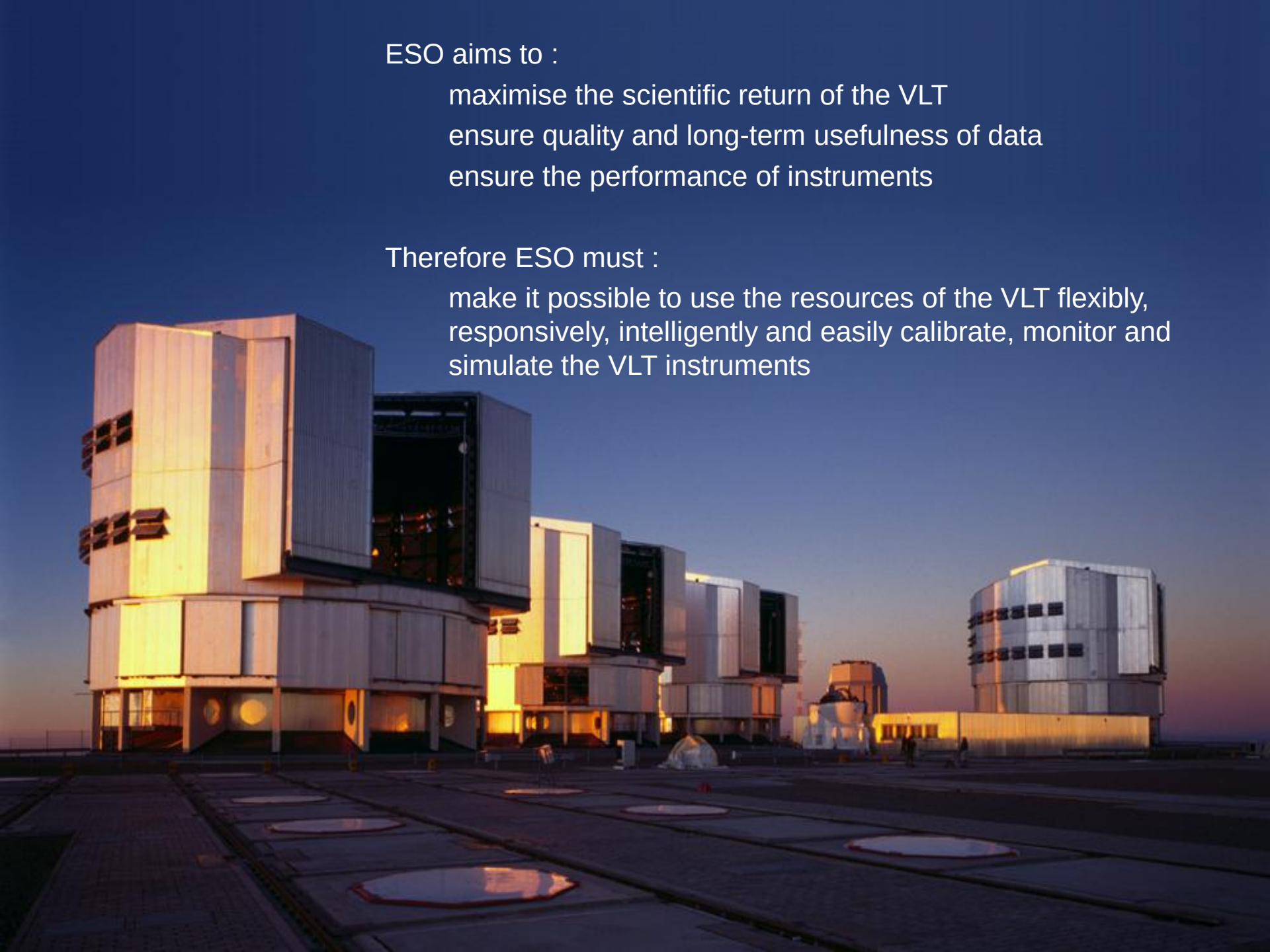
- **La Silla site with medium-sized telescopes**
- **Paranal site with VLT**
- **Strong technology programme with industry and research institutes**

- **Meetings, publications**
- **Large on-line data archives Space Telescope/European Co-ordinating Facility (with ESA)**
- **Student and fellows programmes**
- **Public information and educational programmes**

The Sites

- Paranal
- La Silla
- Santiago
- Chajnantor
- Garching bei München





ESO aims to :

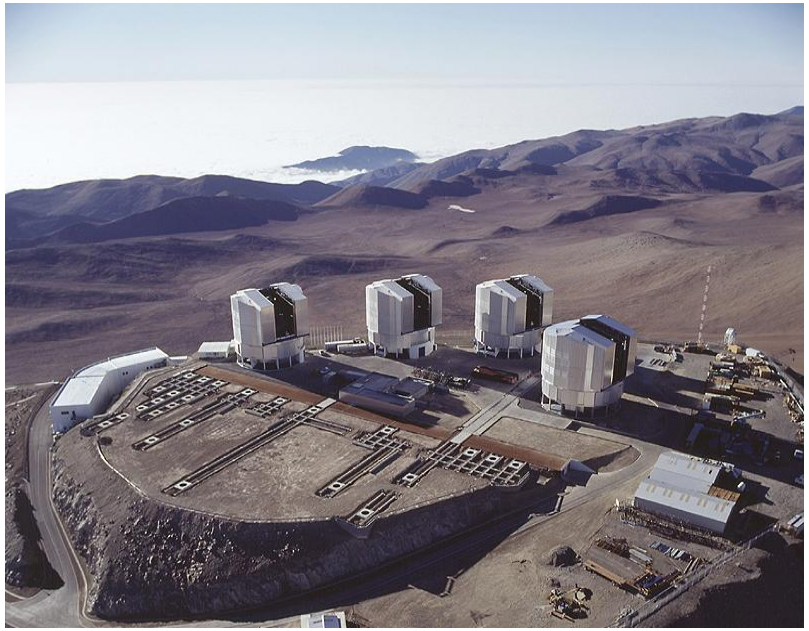
- maximise the scientific return of the VLT
- ensure quality and long-term usefulness of data
- ensure the performance of instruments

Therefore ESO must :

- make it possible to use the resources of the VLT flexibly, responsively, intelligently and easily calibrate, monitor and simulate the VLT instruments

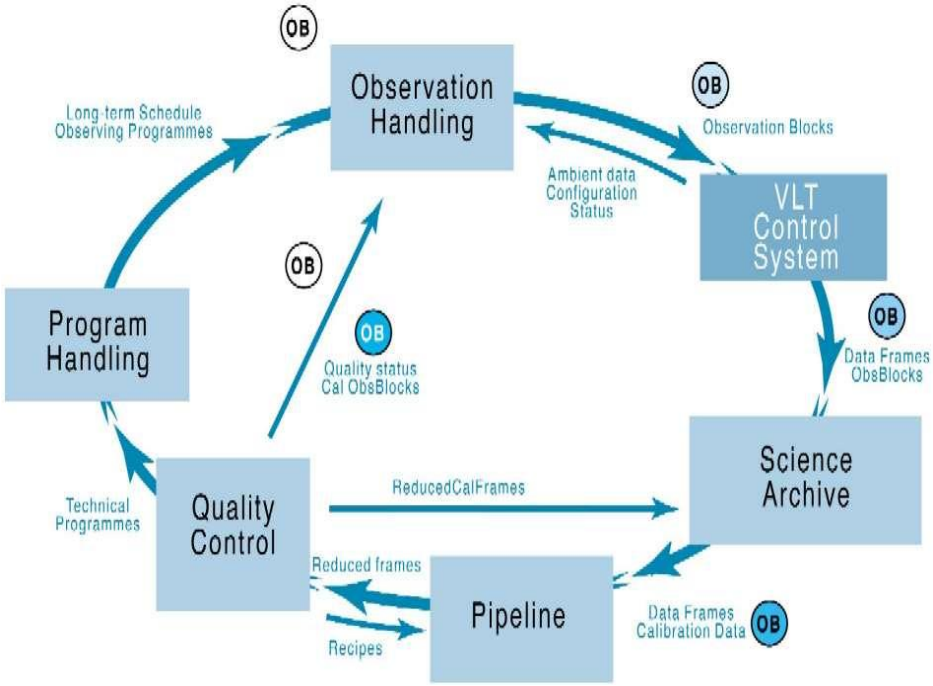
- **To achieve these goals the VLT must have:**

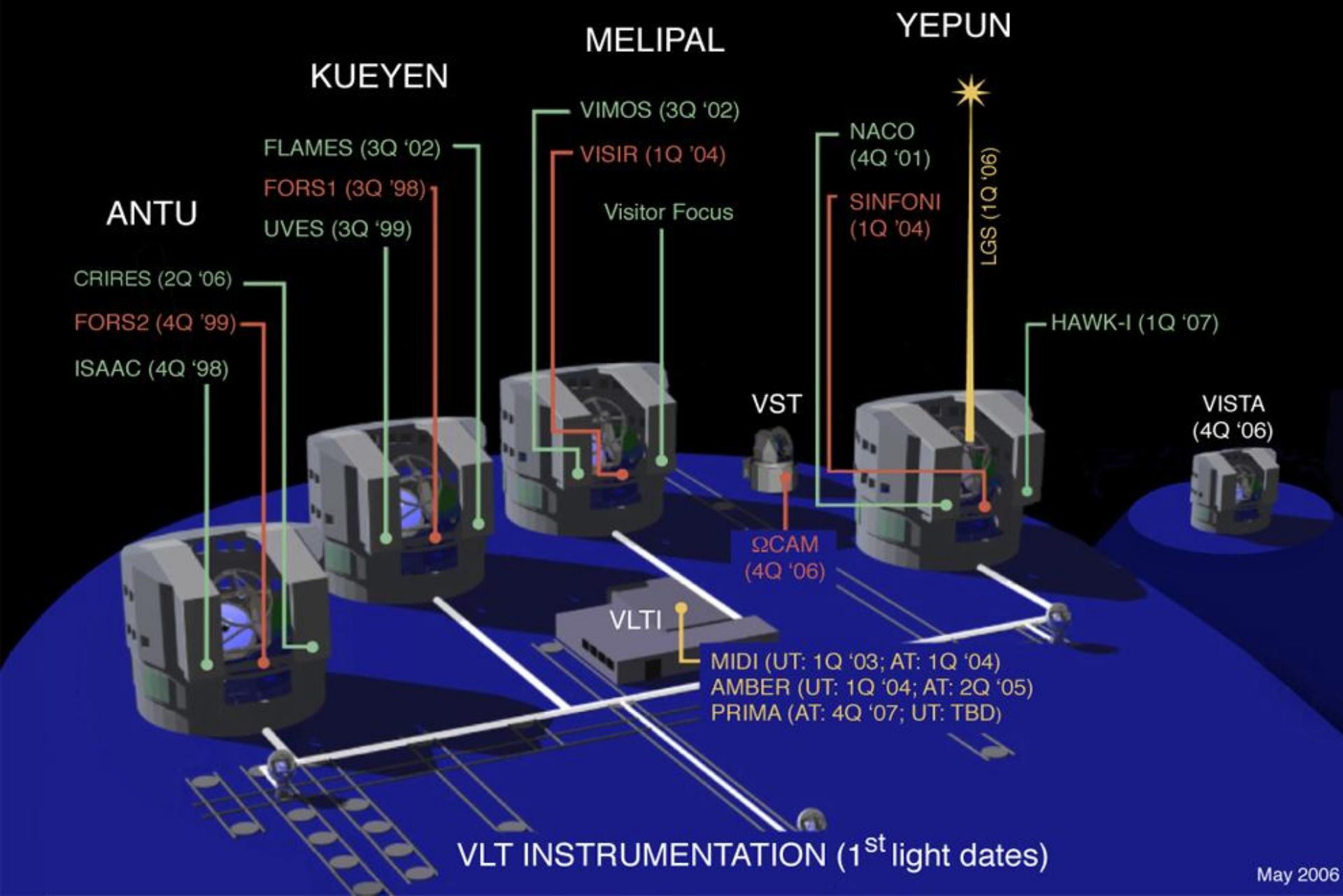
  - a unified, simple interface to telescope and instruments
  - processing on-the-fly
  - flexible scheduling
  - calibration plans for all instruments
  - automatic monitoring of instrument performance
  - detailed instrument models



DFS Mission

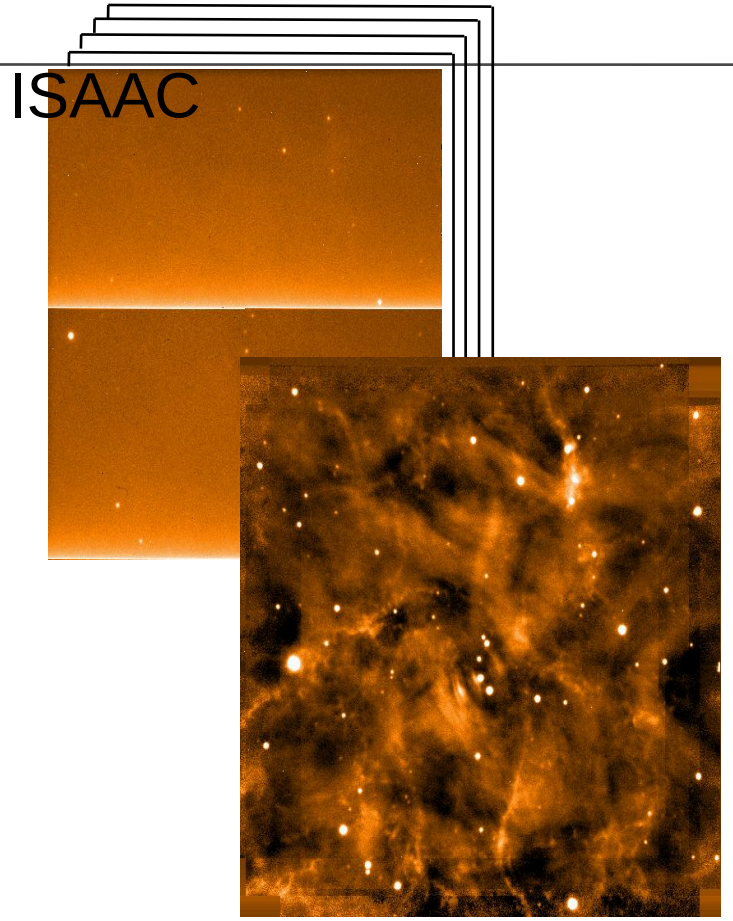
- ***The Data Flow System Department designs, develops and maintains the components of the ESO Data Flow System that are critical to the end to end operation of the VLT, VLTI, VST and VISTA***
- ***Our customers are:***
  - Visiting Astronomers Section
  - User Support Group
  - Data Flow Operations Group
  - Paranal Science Operations
  - La Silla Science Operations
  - ESO Community




MELIPAL
YEPUN
KUEYEN
VIMOS (3Q '02)
VISIR (1Q '04)
FLAMES (3Q '02)
FORS1 (3Q '98)
UVES (3Q '99)
NACO
(4Q '01)
SINFONI
(1Q '04)
LGS (1Q '06)
ANTU
Visitor Focus
CRIRES (2Q '06)
FORS2 (4Q '99)
ISAAC (4Q '98)
HAWK-I (1Q '07)
VST
VISTA
(4Q '06)
ΩCAM
(4Q '06)
VLTI
MIDI (UT: 1Q '03; AT: 1Q '04)
AMBER (UT: 1Q '04; AT: 2Q '05)
PRIMA (AT: 4Q '07; UT: TBD)
VLT INSTRUMENTATION (1st light dates)
May 2006

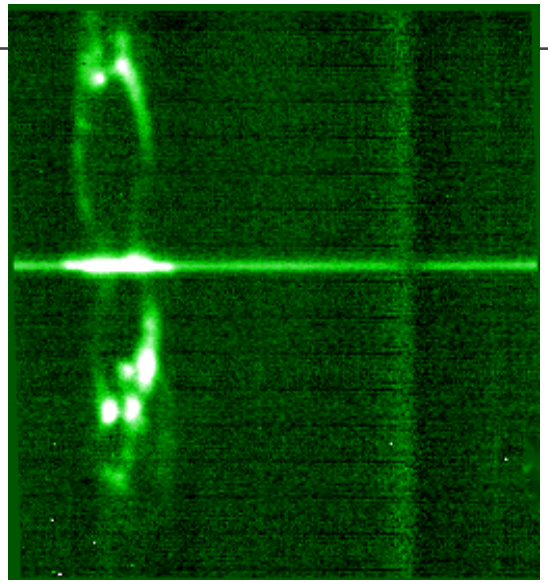
VLT/VLTI Pipelines

- **The main missions of the instrument pipelines are:**
  - Process raw calibration frames into master calibration
  - Produce Quality Control parameters for monitoring telescope, instrument and detector performance
  - Process raw science frames into science data products



VISIR Pipeline

Ant Nebula (Comm III, Sep 2004)



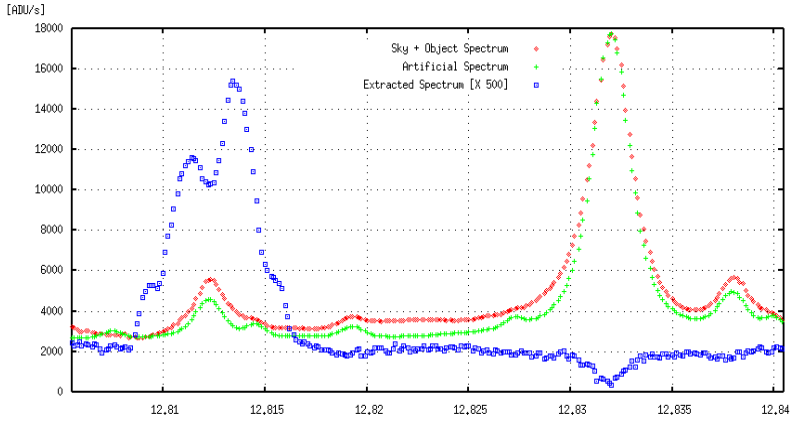
- **Features:**

Imaging

- Chopping/nodding frames reconstruction
- Shift-and-add to create the combined image
- Strehl and sensitivity measurements (STD)

Spectroscopy

- Wavelength calibration
- Spectrum extraction
- Spectral sensitivity measurement



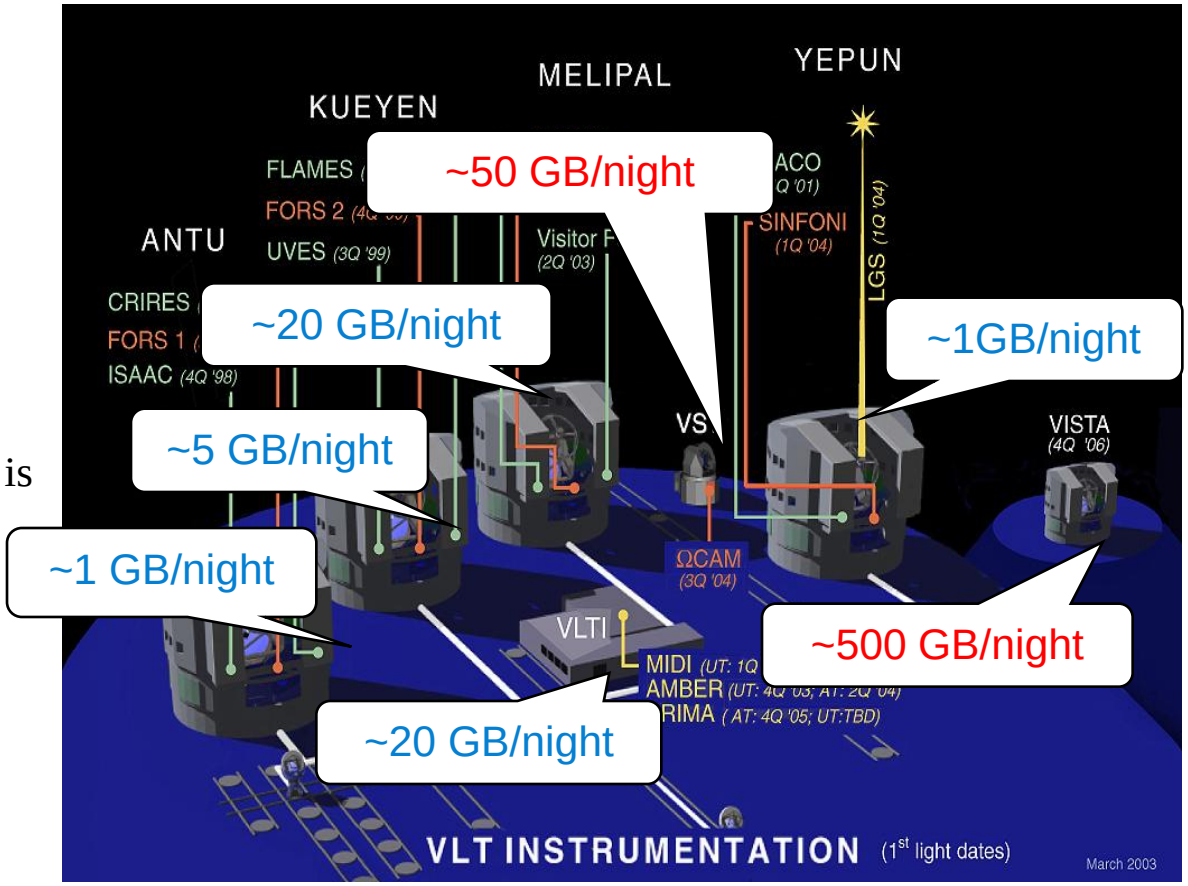
Large Data Rates and Volumes- Challenges

Data Archiving

- Next Generation Archive System

Data Processing

- On-line environment : a new environment based on CONDOR is being developed to replace sequential scheduling of data reduction.
- Desktop environment : Reflex



ESO VLT parallel data processing

- Historically telescope data rate < CPU speed
- Multiple detectors (CCDs) change this
- Most processing is local, i.e. trivially parallel.
- Must avoid I/O bottleneck due to single file for entire detector set.
- Exception is images combined across detectors (e.g. MUSE 2nd generation VLT instrument)