Java and High performance computing

Java and High Performance Computing some experiences from Gaia Data Processing

William O'Mullane

Gaia Science Operations Development Manager

European Space Astronomy Centre (ESAC)

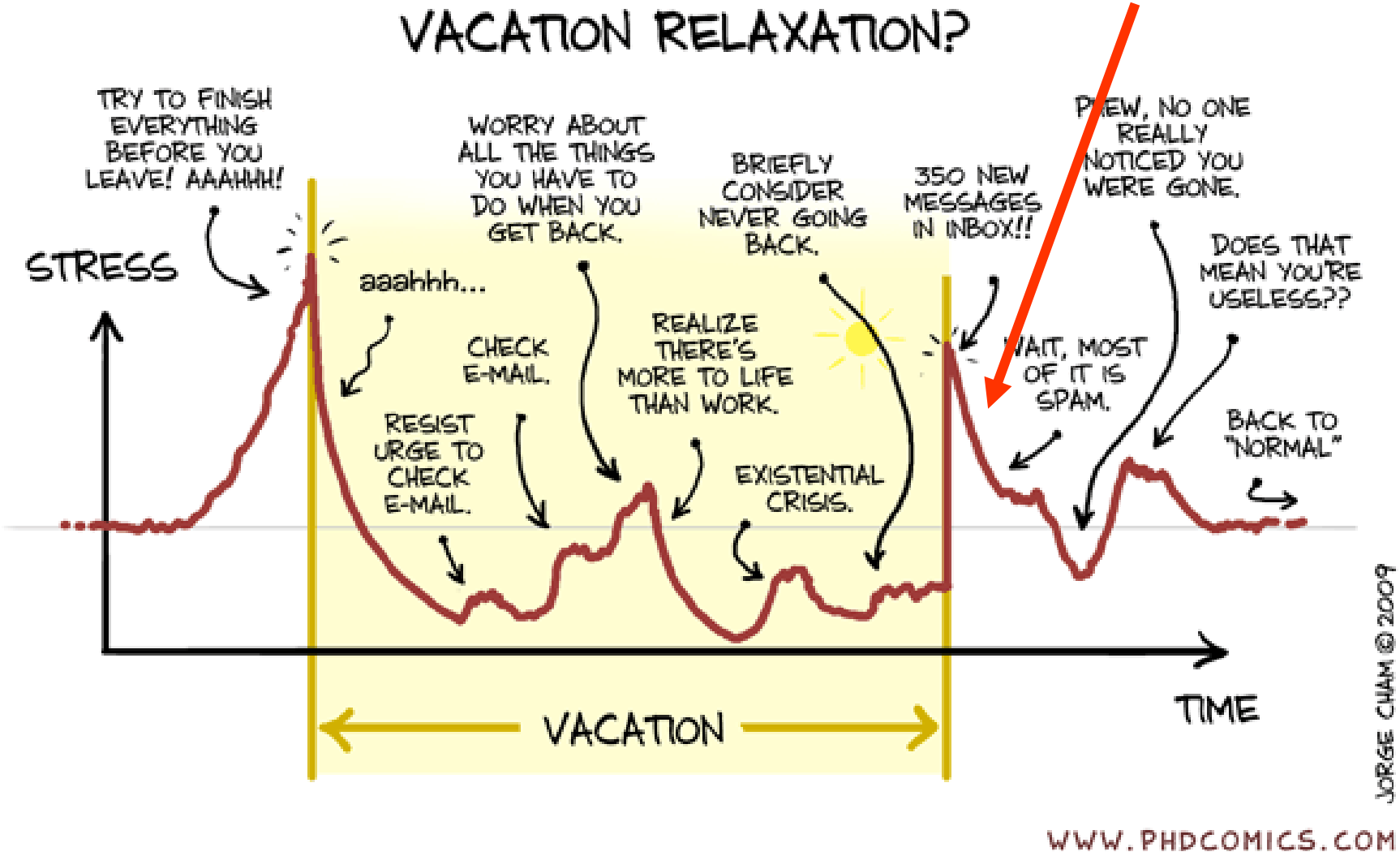
Madrid

Spain

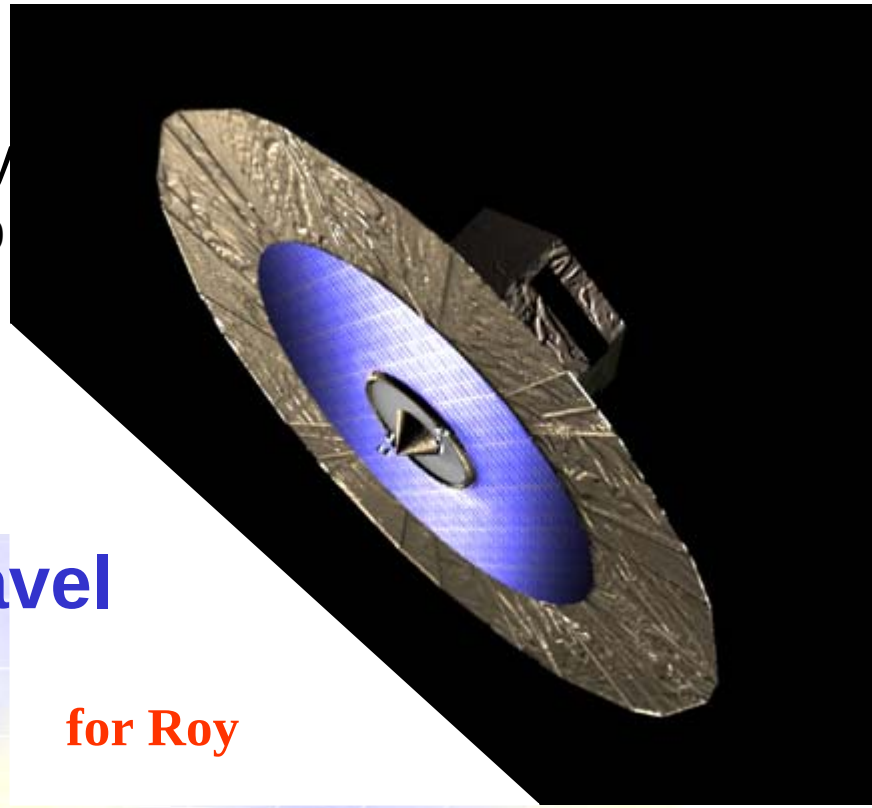
http://www.rssd.esa.int/Gaia

From Dafydd .. while on Holiday

Am about here



Satellite – Obligatory Reminder



- Mission:
  - Stereoscopic Census of Galaxy
  - μarcsec Astrometry G<20 (10^{9} sources)
  - Radial Velocities G<16
  - Photometry G < 20

Discover structure and unravel formation history of Galaxy.

for Roy

- Status:
  - ESA Corner Stone 6
    - ESA provide the hardware and launch
    - Launch: Spring 2012.
  - Satellite In development
    - EADS/Astrium

Launch spring 2012 from Kourou

Gaia is real and getting ready …
Much hardware is ready :
Taurus, some mirrors most CCDs, Thermal tent …

Not to say everything is totally rosy ..
We have several hardware problems like most missions:
CTI effects, PEM non uniformity ..

Image – Soyuz Launch pad under construction
First Soyuz launch from Korou now delayed to April 2010

the topic in hand

- Scientist, Programmer or Manager (or some blend)
- Programmer (Java these days when not managing)
- First line of BASIC when ~14
- Then Machine code and Assembler
- Moved on to Pascal (Borland Turbo !) at 17
- Many hard years with C++ (I have the purify mug)
- Java found me in 1997 while doing come C
- By 1999 never Wanted to look back ..
  - Presented only Java poster at ADASS '99 (I think)
- Also planted first Gaia Java seeds back then
- In 200x Java was the Gaia language of choice
- And then Peter Bunclark† fell off his stool …

Gaia Data processing

- Practically all processing code in Java
  - (one Verification system is not Java)
  - Coding in earnest since 2006
- Simulations started in Java 1998
  - Now on Marenostrum, ranks in top 10-50 supercomputers
  - So we do use Super computers
  - Several TBs of data simulated, transferred
    - Some of it was even processed ..
- Typically highly distributed :
  - No MPI libraries used
  - No Grid libraries used
  - Java has enough functionality for what we need
- Have single data model supported by web based collaborative tool (Hernandez)
  - Generates Java, DB Schemas, even docs in Latex

Is Java fast enough..

- On some processors with highly tuned C compilers the C can be faster than java (max factor 2)
  - You can play cat and mouse for ever with any specific piece of code ..
- On most Intel's Java is as fast or faster than C
  - JIT(Just In Time) Compiler with Hotspot remarkable!
- Just one example from Gaia
  - Relativity C code running in simulator 10 years on super computer
  - Author rewrote in java
  - Its is ~10 times faster in JAVA !
- See Joliet Poster for kind of thing which is much more difficult in C/C++ but kind of natural in Java

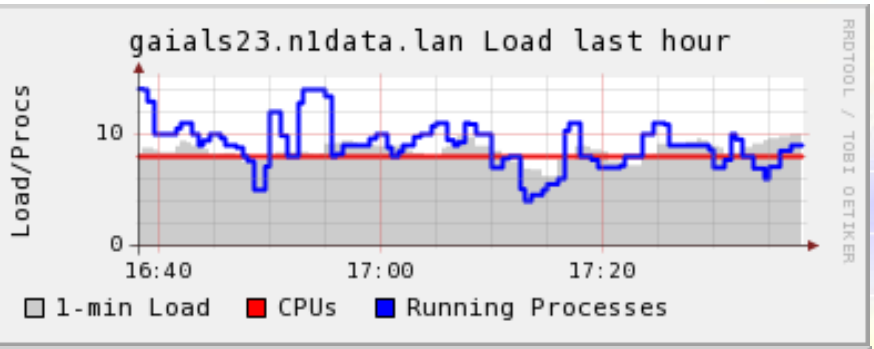
We are in the large systems section

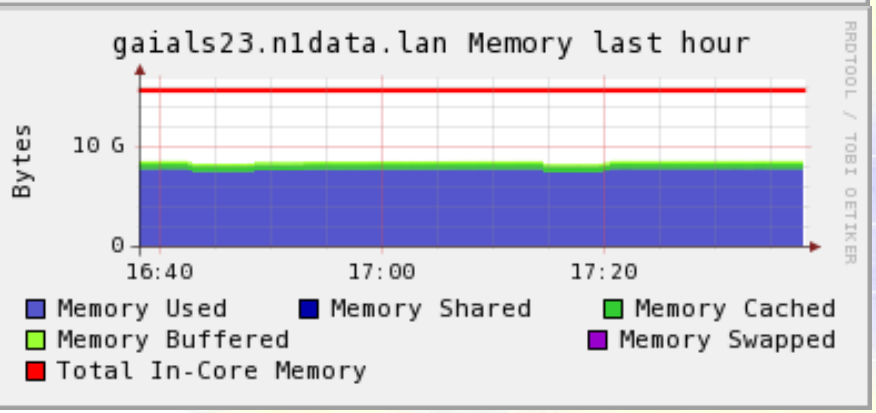
- All metrics are pretty bad but here are a few
- > 10K Person Years effort estimate to build
- Currently > 360 active in dev (>10% FTE)
- Now have >1 million lines of Java
  - estimate is 3 to 5 million lines by launch… in 2020 – who knows !
- Compare (http://en.wikipedia.org/wiki/Source_lines_of_code)
  - Windows NT ~40 million
  - Linux kernel ~ 5 million
  - Debian >200 million

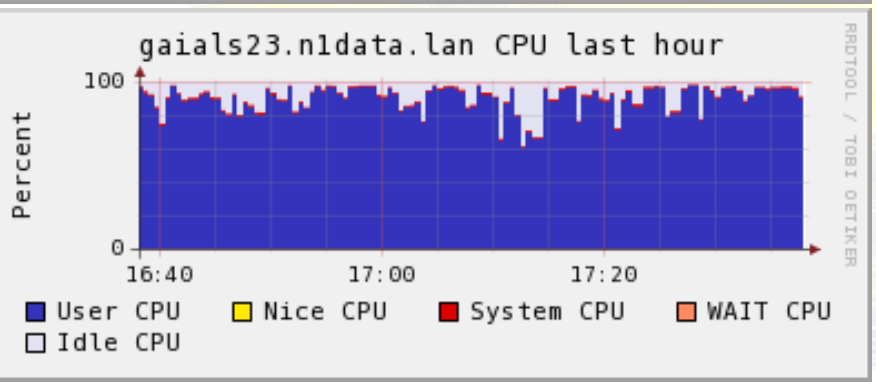
And what is HPC

- High Performance Computing –
  - No definitive definition ..!
  - Wikipedia of course - "**High-performance computing (HPC)** uses supercomputers and computer clusters to solve advanced computation problems. "
  - Pretty broad .. Usually also taken to mean getting most out of machine
  - These days power consumption also on the agenda

Efficiency – Crux of the issue







AGIS runs like this

High load, low network, high CPU!

Some HPC people tell us we should rewrite in C (save energy)

- Next slide

Supercomputer centres seem to have very specific macros to include in C code to make it efficient for THEIR machine.

- looks a little like a virtual machine
- Why not provide a better JVM for their machine ?
- Or windows CLR ?

(Re) code for speed – save energy cost.

- How often would I Have to recode (each evolution)
  - Gaia coding started 2006 (or earlier)
  - Will end 2017 or 2020
  - How many evolutions will a Supercomputer go through …
- just to code in C may give fact 2 speed (with Cost)
  - Past effort + future = ~30PY IN JAVA + maintenance 7PY
  - In C we estimate this to be ~Double ~70PY
  - So just CODING in C might cost ~3.5MEuro MORE
    - Even if (ab)used postdocs it is a lot …
  - Energy for running AGIS (10^20 FLOP) ~150KEuro
    - Ok with idle time and increasing energy costs perhaps 1MEuro
- recoding … well lets not even go there Huge Uncertainty

Java = Portability

- Java = Portability right !
- Well - Yes it does
- Ok some problems with Graphics
  - mostly we are not doing that
- This is a HUGE advantage
- This big cloud just rolled in
- We could use that right ……

AGIS on the cloud

- See Lammers POSTER for more on AGIS
- Took one person less than one week to get running (Parsons,Olias).
  - Main problem DB config
  - Also found scalability problem in our code (never had one hundred nodes before)
- It ran at similar performance to our in house cheap cluster.
  - E2C indeed is no super computer
  - Oracle image was available already
  - AGIS image was straightforward to construct but was time consuming – better get it correct !
- Availability of large number of nodes very interesting –not affordable in house.

Cost effectiveness of E2C.

- AGIS runs intermittently with growing Data volume.
- Estimate 2015 ~1.1MEuro (machine) + 1Meuro (energy bill less ?) = ~2Meuro
  - In fact staggered spending for machines
  - buy machines as data volume increase
- Estimate on Amazon at today prices with 10 intermittent runs ~400Keuro
  - Possibility to use more nodes and finish faster !
- Reckon you still need in house machine to avoid wasting time testing on E2C
- Old nut, Vendor lock-in ? Need standards

Conclusion

- Many projects now using Java
  - ESA - Herschel, Plank (a little) , archives
  - Saw some in the Spitzer slides
  - JWST (also C++..)
  - Lots of others and probably more to come
- Others using other HIGHER level languages
  - SkyServer from JHU/SDSS – C# .NET
  - JWST again have some C# and lots of Python
- Portability maintainability as important as speed for processing systems
- So we are already doing it (HPC or not)
  - “Yes we can”

Questions ??

Ariane V188 carrying Herschel and Planck (May 14 2009)