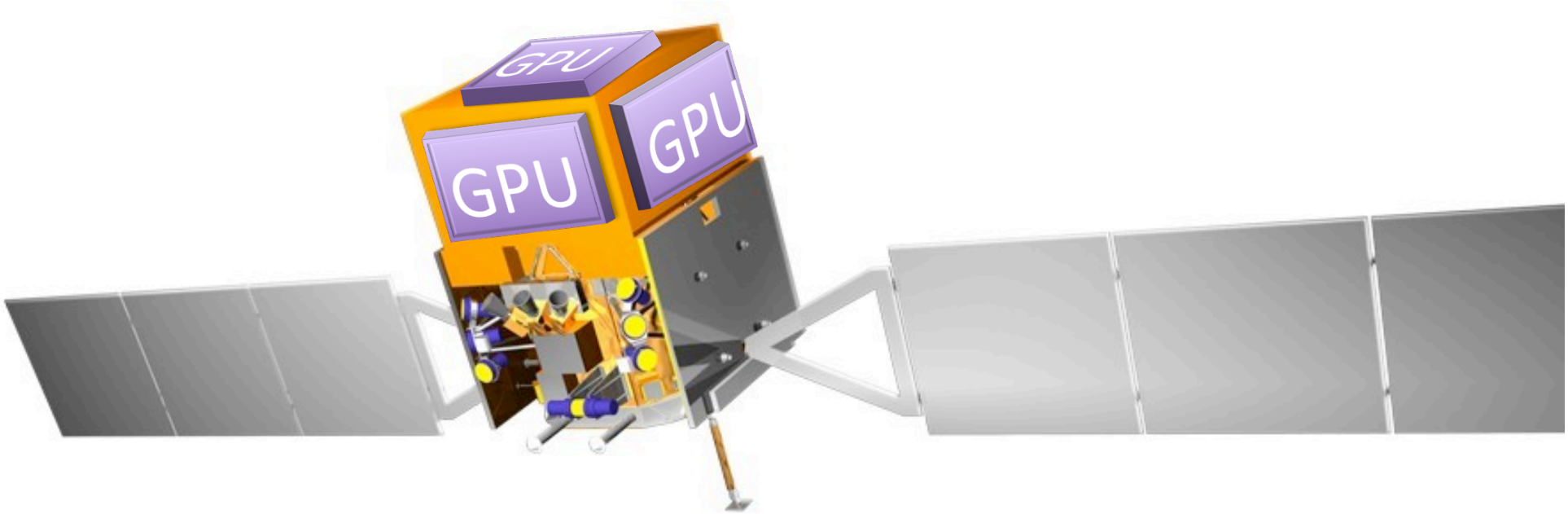
Porting Fermi Analysis Tools to GPUs



Denis Bastieri
INFN/Univ. Padova

Scientific Computing Workshop, 20 July 2011, SLAC

LAT Analysis: overview

- Standard Maximum Likelihood approach:
  - Create a model
  - Compute the likelihood of your model with respect to the detected gamma rays
- Steps (tools developed by Jim Chiang):
  - Selection: data stored in FITS format. *gtselect*: apply required selection
  - Evaluate likelihood: could be factorized.

LAT Analysis: steps

First trial: unbinned likelihood

0. prepare the (XML) model
1. gtselect: apply desired cuts
2. gtltcube: compute the *livetime cube*.
3. gtexpmap: compute the *exposure map*.
4. gtlike: compute the likelihood

versatile |ˈvəːsətʌɪl|

adjective

1 able to adapt or be adapted to many different functions or activities *: a versatile sewing machine | he was versatile enough to play either position.*

2 archaic changeable; inconstant.

DERIVATIVES

versatilely adverb

versatility |-ˈtɪlɪti| noun

ORIGIN early 17th cent. (in the sense [inconstant, fluctuating]): from French, or from Latin ***versatilis***, from ***versat- 'turned around, revolved,'*** from the verb ***versare***, frequentative of ***vertere 'to turn.'***

Where?

home people hardware projects contacts

mimesis

simulation and scientific computing

Computing facilities

Urania 2 x Xeon E5620 2 x NVIDIA S2050 3584 GPU cores - 24 Gb RAM

... men who have been instructed of her she raises aloft to heaven (ouranos), for it is a fact that imagination and the power of thought lift men's souls to heavenly heights ...

Clio 2 x Xeon E5620 4 x NVIDIA GTX 580 1920 GPU cores - 12 Gb RAM

... the praise which poets sing in their encomia bestows great glory (kleos) upon those who are praised ...

Euterpe 2 x Xeon E5620 1 x NVIDIA S2050 1792 GPU cores - 24 Gb RAM

... she gives to those who hear her sing delight (terpein) in the blessings which education bestows ...

gtltcube: the easy part

- Fixed dimension *ltcube*:

```
#define CUBE_DIM_RA 100
#define CUBE_DIM_DEC 100
#define CUBE_DIM_COSTHETA 100
```

- 1 year (= $\pi \times 10^7$ s) / 30 s: 1 Mrow
- Investigating with HEALPix 1° (42k),
  but dcostheta = 0.025
- Everything stored in GPU memory
  (not a big gain: many operations on each row)

gtltcube: implementation

- kernel: loop on the seconds in the 30s interval
  - find the zenith
  - loop on θ
  - loop on ϕ (0 .. 360×sinθ)
  - update matrix(θ) **!!avoid conflict on GPU!!**
- CPU: loop on the Mrow
- GPU: launch 1 Mthread
- 80× faster on GPU!

gtexpmap:
adding diffuse sources

- The *ltcube* is the input of the gtexpmap.
- My *ltcube* overly simplified: ra/dec -> HEALPix
- lat/long matrix: typical binning 0.5°

  ⇒ 25° radius: nlong=100, nlat=100
- Energy binning: 4 or 5 bins/decade will do

  ⇒ 20 bins in energy (usually 100 MeV – 300 GeV)
- More accuracy for diff.srcs? More lat/long bins!

  ⇒ Test a set of optimizations “QVA”

The sw/hw architecture

1. Allot one GPU card for ltcube: FT2 (spacecraft) data resident in memory, output to…
2. … a (different) GPU card allotted for expmap: FT2 data resident in memory, output ltcube and expmap to…
3. … a third card storing FT1 data, where gpulike will select data and compute the likelihood of the models fed.

DB Wrap™

- gtselect: 10-yr mission = 2 Gγ = 200 GB !!
- New in Cuda 4.0: unified access:
  40 bit => 1 TB
- **DB Wrap™**: SQL-complying, DB-agnostic, middle-layer, GPU<->DB interface.
- Performs SQL-select launching a kernel on each selected entry

DB Wrap: implementation

- 1 TB too expensive & not needed for LAT
- Reduce row size (time, ra, dec, energy, quality)

  ⇒ 92 B/entry -> 20 B/entry -> 4GB/year
- Select only good quality data (0.1×)

  ⇒ (20 GB)/(10 year) fits in 6 (or 4) GPU cards
- Everything driven by a single host!

Next steps

- ltcube and expmap server ?
- Choosing a proper hw architecture:
  3 or 4 GPU/PCI Express ? , 2 PCI Express/host !
- Dimensioning γ data accordingly
- DB Wrap: still few issues to solve
- QVA: is it really needed ?
- HEALPix fns in a kernel: thread-safe ?

Welcome to real-time analysis!