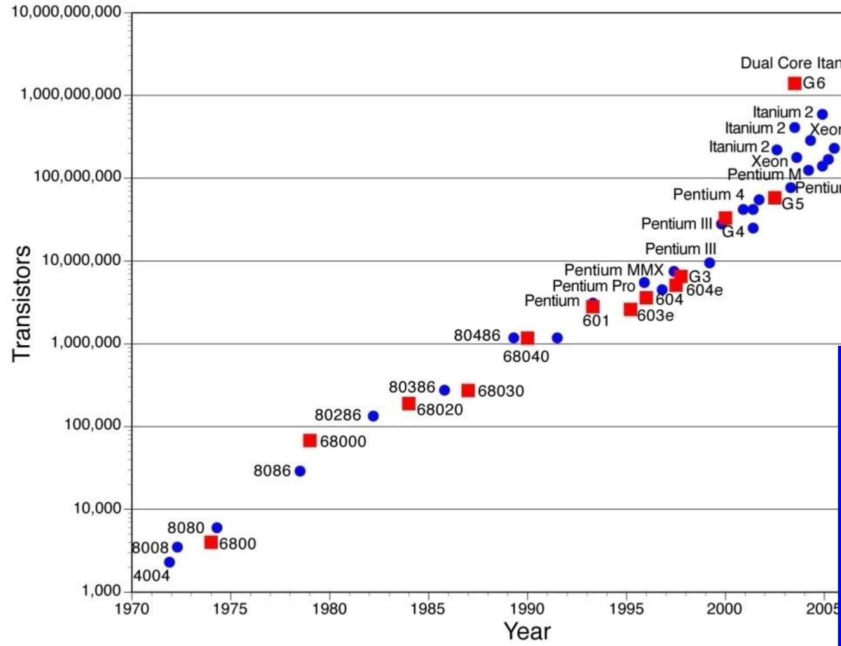
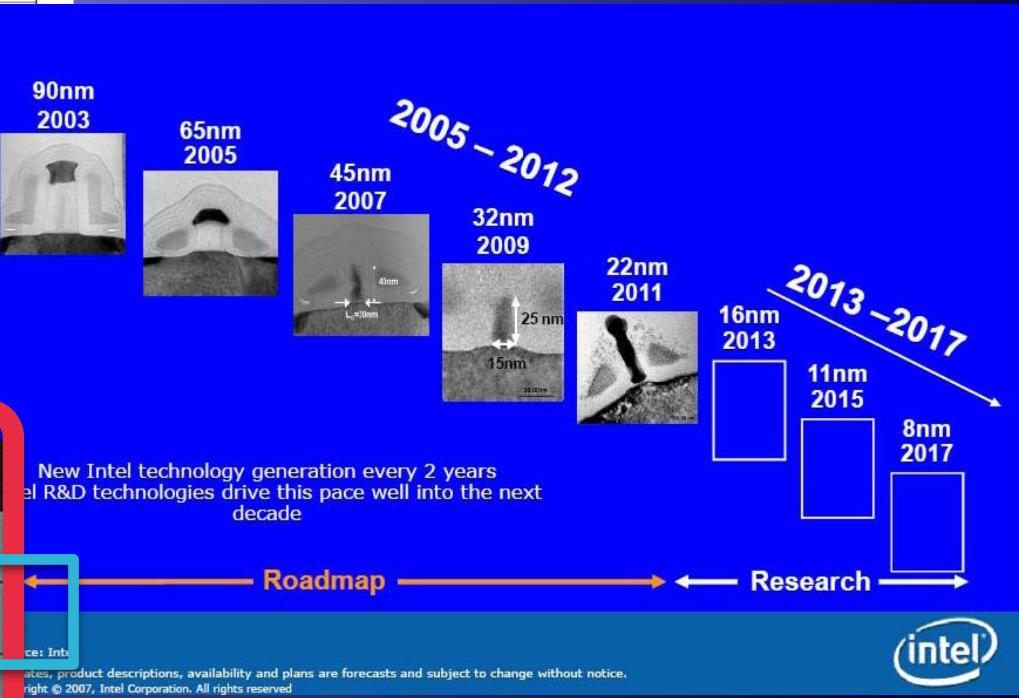
# Computing infrastructure manycore architectures perspective

Dr. Leone B. Bosi – INFN Perugia

ET – WG4 Meeting
2 October 2009

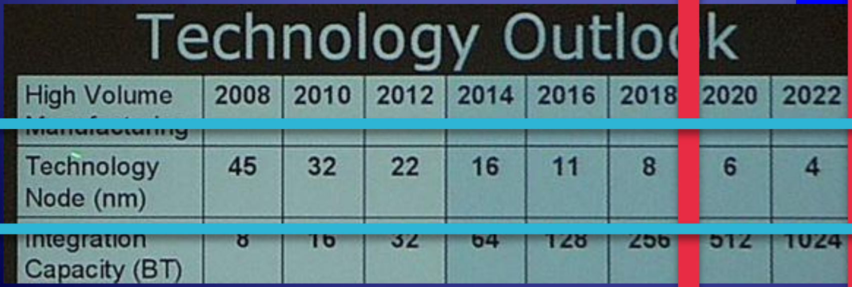
# Technological outlook





New Intel technology generation every 2 years
el R&D technologies drive this pace well into the next decade

ce: Inte
ates, product descriptions, availability and plans are forecasts and subject to change without notice.
right © 2007, Intel Corporation. All rights reserved



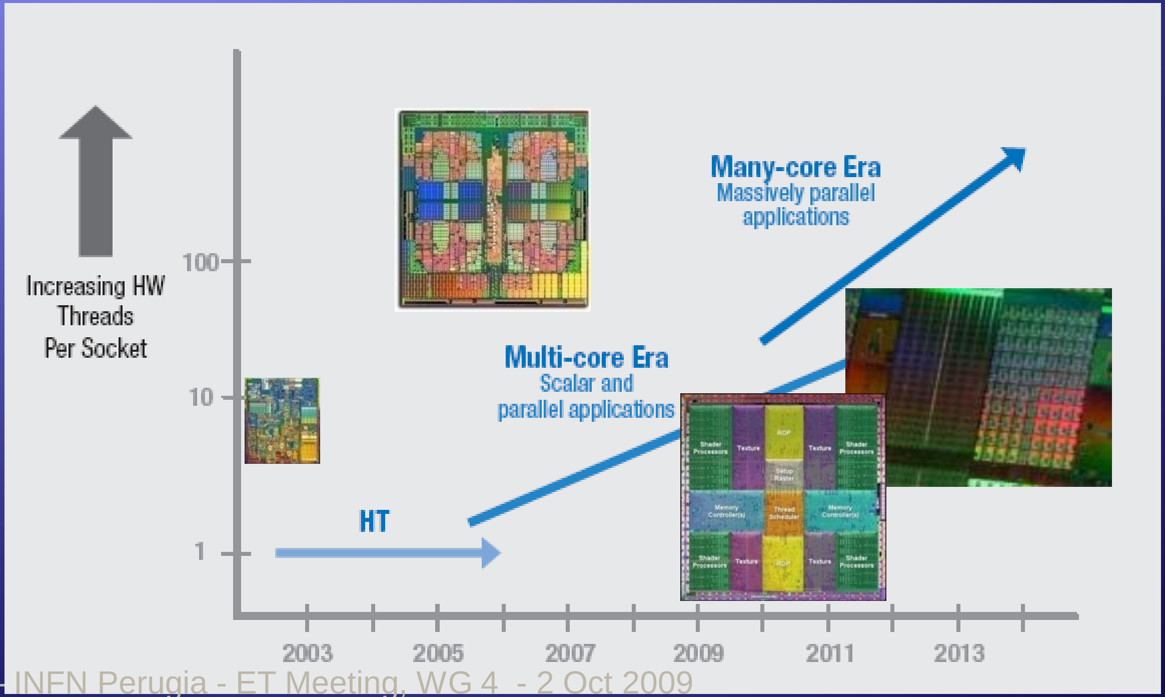
## Technology Outlook

| High Volume Manufacturing | 2008 | 2010 | 2012 | 2014 | 2016 | 2018 | 2020 | 2022 |
|---|---|---|---|---|---|---|---|---|
| Technology Node (nm) | 45 | 32 | 22 | 16 | 11 | 8 | 6 | 4 |
| Integration Capacity (BT) | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

Intel ref.

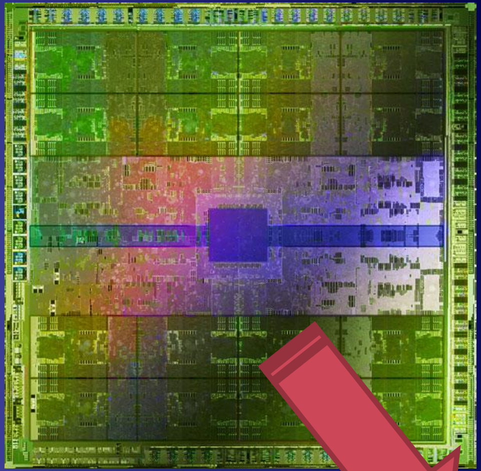
# Technological outlook

- Most important chip semiconductor maker are working in order to limit problems due to integration scale reduction.
- In fact last 10 years the processors architectures are changed a lot, introducing parallelization at several architectural levels.
- That evolutive process will continue in a ever more deeper manner, moving to the so called “many-core” era.



Intel ref.

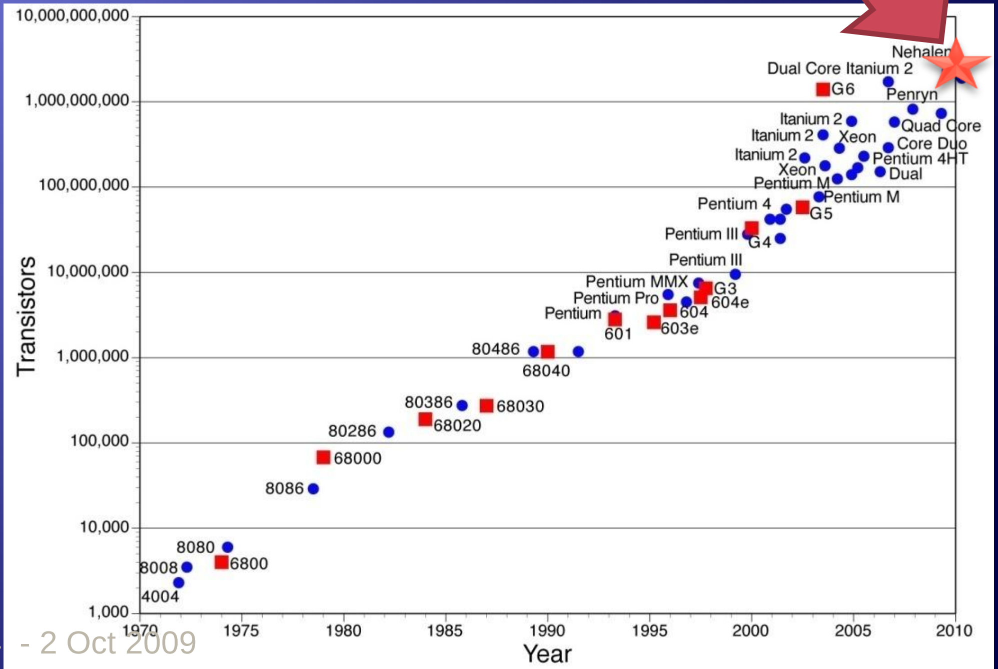
# Status of art



Nvidia Fermi core:

- The GPU is made up of **3.0 billion transistors** with 40nm technology.
- **512** CUDA processing cores organized into 16 streaming multiprocessors of 32 cores each.

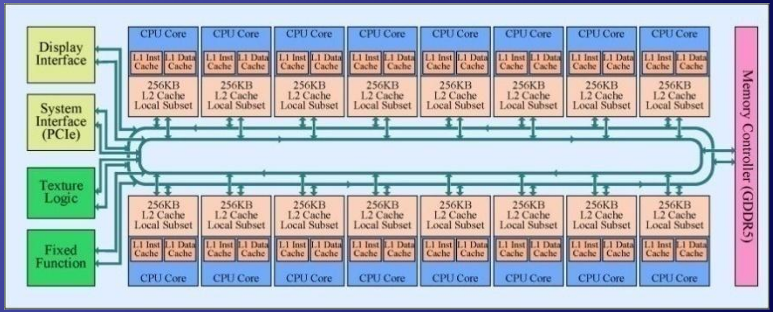
- The memory architecture is built around a new **GDDR5**
- **six channels of <u>64-bits</u>** for a total memory bus of **384-bits**.
- The memory system can technically support **up to 6GB** of memory



# Status of art:

**AMD ATI Radeon HD 5850**

- **2.15 billion 40nm transistors**
- **1440** Stream Processing Units
- **1600** shader units, divided in **20 core SIMD** with **16 stream** processor each one
- Declared Peak power:
  - Single precision: **2,7 TFLOPs**
  - Double precision: **544 GFLOPs**



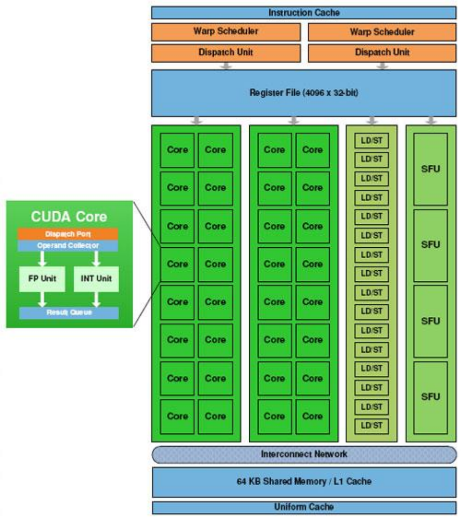
**Intel Larrabee**

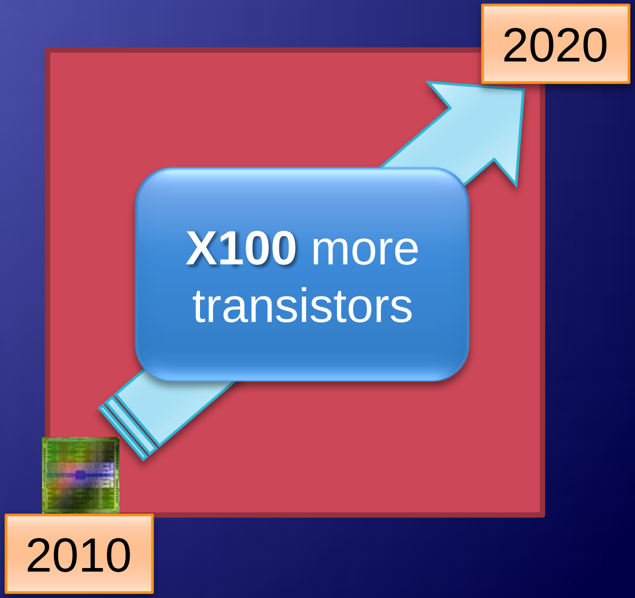
- **2 billion 40nm transistor**
- 32-48 SIMD core x86-64, based on Pentium P54C
- **Each core** 512-bit vector processing unit

  =16 precision floating point numbers at a time

# Some performance considerations:

- These new architectures require a complete different programming models.
- In future, in the manycore era, computing power will be distributed across multiple cores (10000>?) on a single processor, and many processors on a single board.
- Performance achievable from these architecture is not predictable because depends on algorithm and relation with:
  - Memory/registries architecture model
  - Intercommunication
  - Serial portion of the algorithm





2020

**X100** more transistors

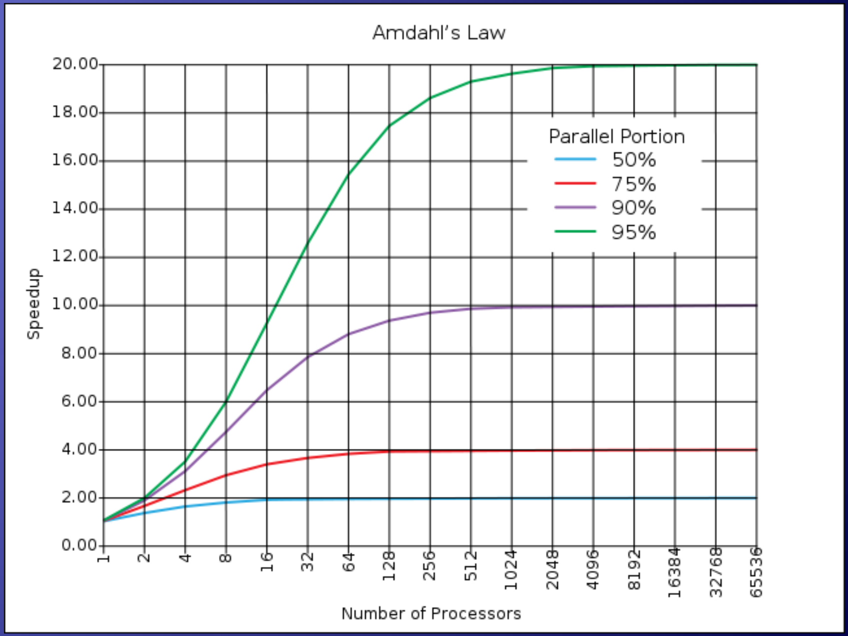
2010

# A note on Amdahl’s law:

- The speedup of a program using multiple processors in parallel is limited by the time needed for the sequential fraction of the program.

- The maximum speedup that can be achieved by using *N* processors is: $\frac{1}{(1-P)+\frac{P}{N}}$.

  *where* P is the proportion of a program that can be made parallel.



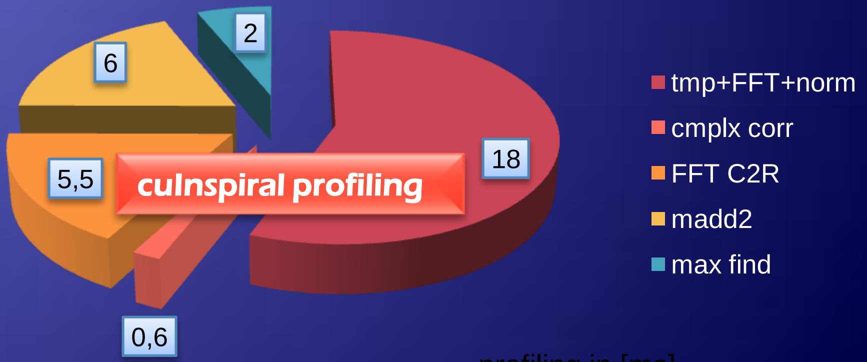
# GPU performances: The CB case (1):

- We have developed **cuInspiral** library a prototype of a full GPU pipeline, permitting to evaluate perspective and potentiality of these new architecture(details will be presented on Erice talk)
- This library implements fully GPU functions (e.g.):
  - Taylor PN2 generator
  - Normalization
  - Matched filtering
  - Maximum identification
- Main performances speedup measured:
  - Template Generation: x100
  - FFT: x60
  - Reduction: x80

# GPU performances: The CB case (2):

- If we consider analysis parameters of :
  - low.cutof.freq:24Hz,
  - vector length 2^20, fs=4kHz the
- The processing rate if roughly of **35 templates/sec** (lower limit)
- The online constrain processing  if of 4000 templates. That means that the Virgo matched filtering only analysis can be performed with a couple of GPU 275 (=500 Euro)

Pipeline Gain (lower limit):
**50** with GTX 275
→ expected  with Fermi GPU: **150**



profiling in [ms]

# Forecast by 2020:

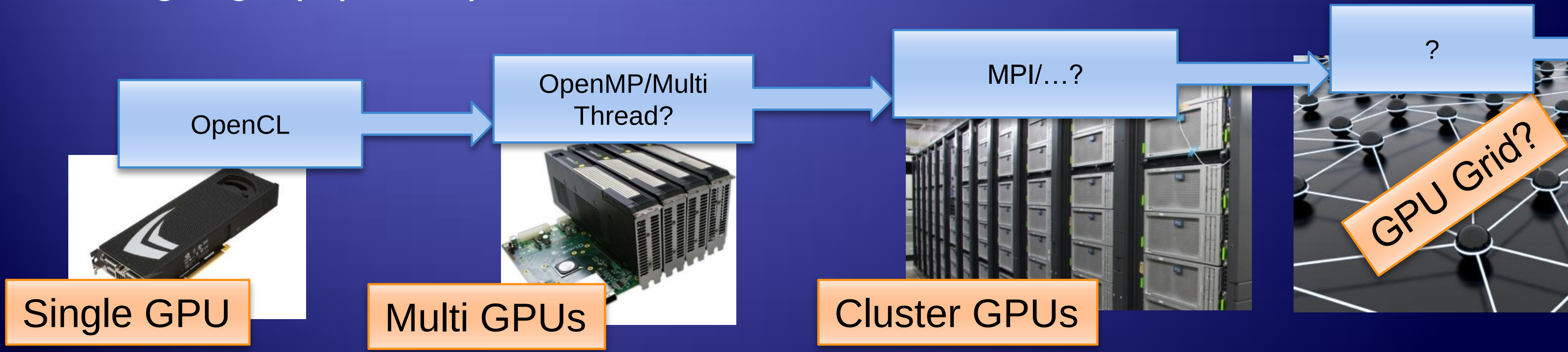
We could try to make a projection of the available computing power by 2020, in the context of CB like algorithms, making some assumptions:

1. We can start considering that the actual firsts attempts of manycore architecture provides a factor **x150** in single precision respect CPU implementation.
2. We can consider a Moore's law factor of **x100**
3. From the experiences coming from massive parallel architecture, usually performances are reduced significantly by communication overhead, thus we take **x0.4** (it could be even worse)
4. We obtain :

- ❑ a gain of a factor **6000** respect the actual CPU implementation.
- ❑ Equivalent to **5 TFLOPs or higher** on a single manycore processor by 2020.

# GPU computing e programming paradigms

- The architectural differences between GPU and CPU are evident, in particular the way how the relations between cores, memory, shared memory and IO subsystem are organized
- Moreover different chip producers implement different solutions with different characteristics and instructions sets
- Recently important efforts have been done  by Apple, Intel, NVIDIA , Sony, … in the direction of programming standardization for parallel architecture: The Khronos Group, defining the Open Computing Language (OpenCL).



OpenCL

OpenMP/Multi Thread?

MPI/…?

?

Single GPU

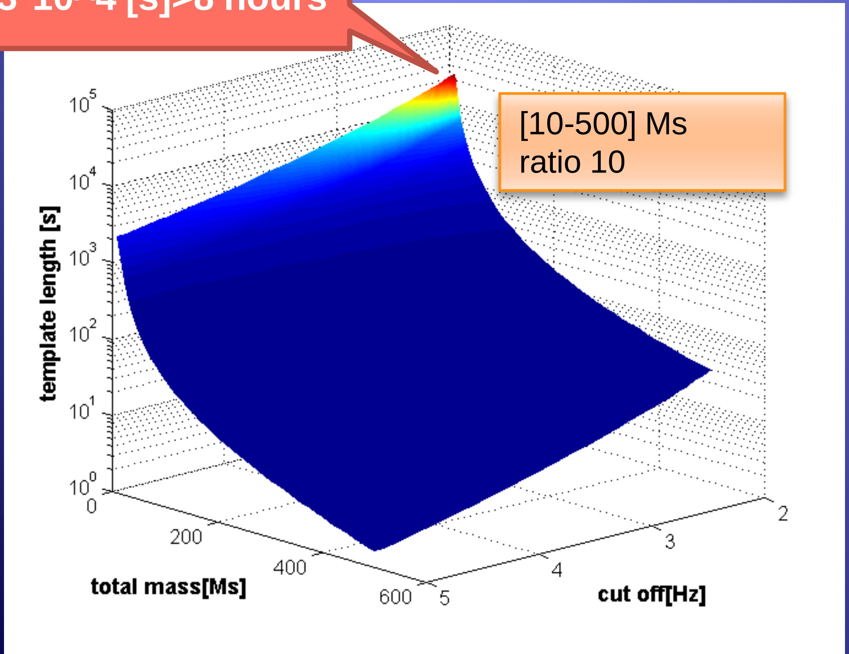
Multi GPUs

Cluster GPUs

GPU Grid?

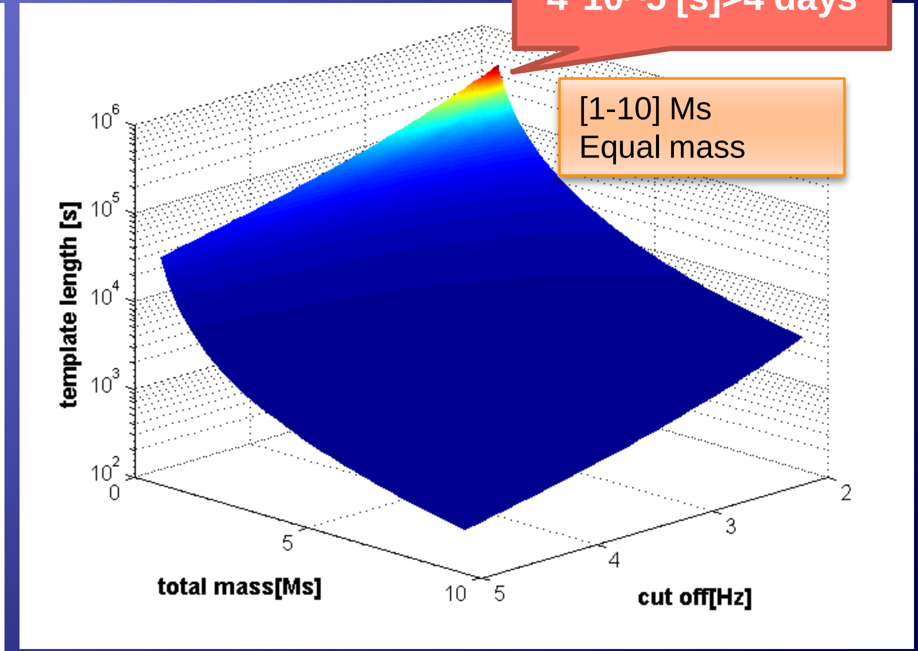
# What with this power? Inspiral case

- ET sensitivity permits to observe Inspiral signals for much more longer time respect actual detectors.



3*10^4 [s]>8 hours

[10-500] Ms
ratio 10

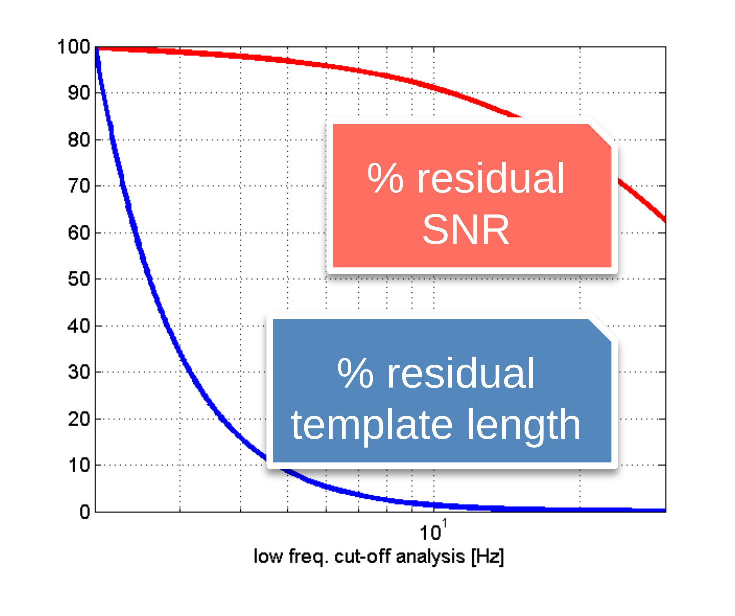


4*10^5 [s]>4 days

[1-10] Ms
Equal mass

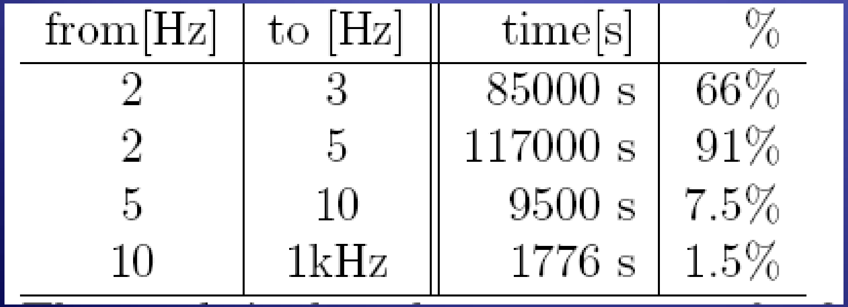
# Inspiral case: data handling

- A template bank roughly estimated for “classical” inspiral analysis is composed by 2000000 templates (mm=95%,[1-500]Ms) (e.g.Virgo @30Hz=7000)
- The complexity gains of about 500 times.
- For longest template we can operate in a more tricky manner→”inverse followup”



% residual SNR

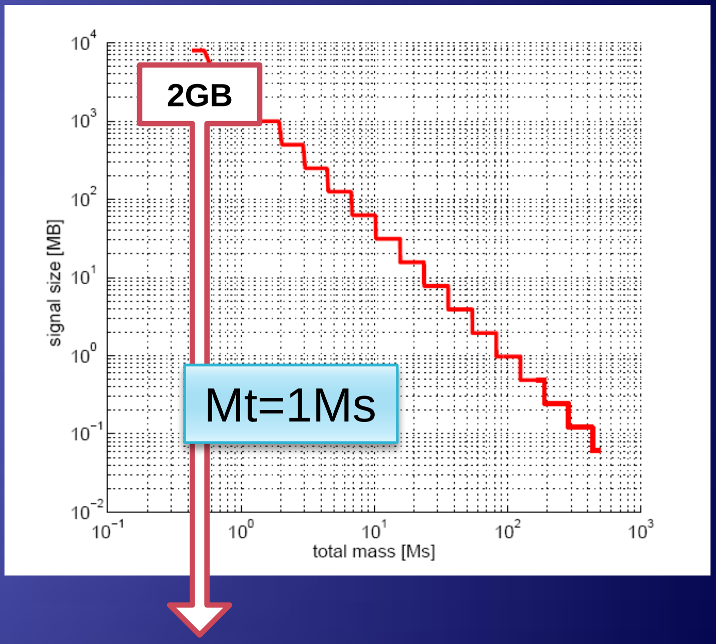
% residual template length

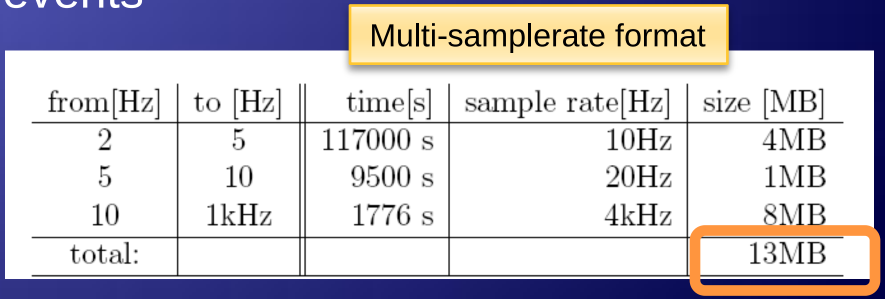


| from[Hz] | to [Hz] | time[s] | % |
|---|---|---|---|
| 2 | 3 | 85000 s | 66% |
| 2 | 5 | 117000 s | 91% |
| 5 | 10 | 9500 s | 7.5% |
| 10 | 1kHz | 1776 s | 1.5% |

# Inspiral case: “inverse followup analysis”

- Divide the investigative process in more steps:
  - The Firsts(detection) try to catch the inspiral process from where it is more important for the detection point of view (@ ET era, templates will be composed by inspiral,merging and ringdown phase.)
  - The next step performs a reverse followup of the events with a multisample rate analysis, introducing more accuracy and following the events evolution
  - switching in observation mode.




Multi-samplerate format

| from[Hz] | to [Hz] | time[s] | sample rate[Hz] | size [MB] |
|---|---|---|---|---|
| 2 | 5 | 117000 s | 10Hz | 4MB |
| 5 | 10 | 9500 s | 20Hz | 1MB |
| 10 | 1kHz | 1776 s | 4kHz | 8MB |
| total: | | | | 13MB |

# Inspiral case: detection computing cost

- Given a template bank for ET, We can define to truncate critical long inspiral by chosing properly the low frequency cut-off (e.g.4Hz)
- With this choice we can reduce the max template length for this analysis step → 3600s
- **2000000** templates (length=2hours@4kHz), today using a culnspiral GPU like library to process a single timeslice we require: 2000000 x 0.8 s=**18 days**!!
- If we renormalize respect the previous estimated gain factor (@2020 forecast) **3x100*0.4=120** we obtain

  →4hours computing time
- It seems plausible that by 2020 computing innovation we will be able to pursue ET requirements for this task.