XFEL offline computing@DESY

Martin Gasthuber, Volker Gülzow

Outline

HAMBURG • ZEUTHEN

- **todays resources and their usage**
- **potential of todays systems**
- **example usage for (known) XFEL demands**
- **more demanding/critical questions**

Part of the analysis chain

HAMBURG • ZEUTHEN

It will be very different from today (requirements)

What are the components?

1) **Network**
2) **Compute Resources**
3) **Storage**
4) **Software**
5) **Support**

-> A computing model/TDR is needed for offline computing!

Part of the analysis chain

HAMBURG • ZEUTHEN

1) **Network: technically okay, bunch of 10/40/100 GigaBit/s Ethernet (GE) connections (money)**

4) **Software: both, technical (OS, Compilers,..) and analysis packages (okay?), Performance? Multicore enabled, scalable?**

5) **Support: Staff for operations/technical SW and staff for analysis packages**

2. Computing Resources

HAMBURG • ZEUTHEN

- **Model is application driven**
- **Expecting all Unix (Linux) driven**
- **Major work on local clusters (farms with fast interconnect eg backplane, infiniband)**
- **Grid solutions**
- **Special system for visualization**

FFT benchmarks

By Frank Schluenzen (IT)

HAMBURG • ZEUTHEN

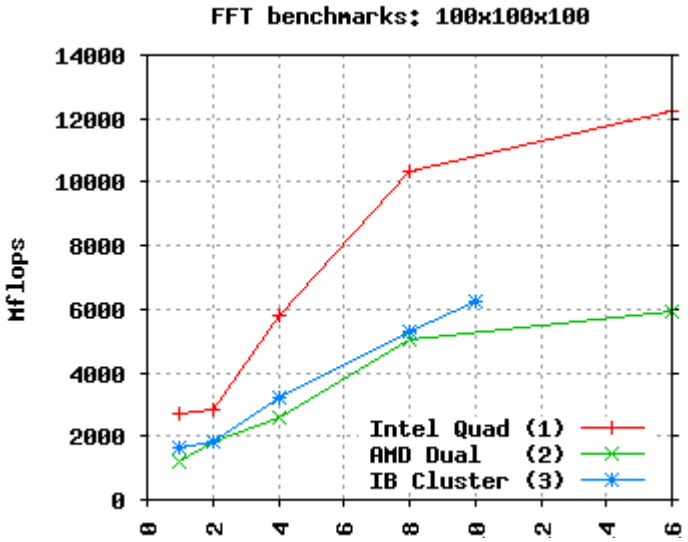
- Available hosts:

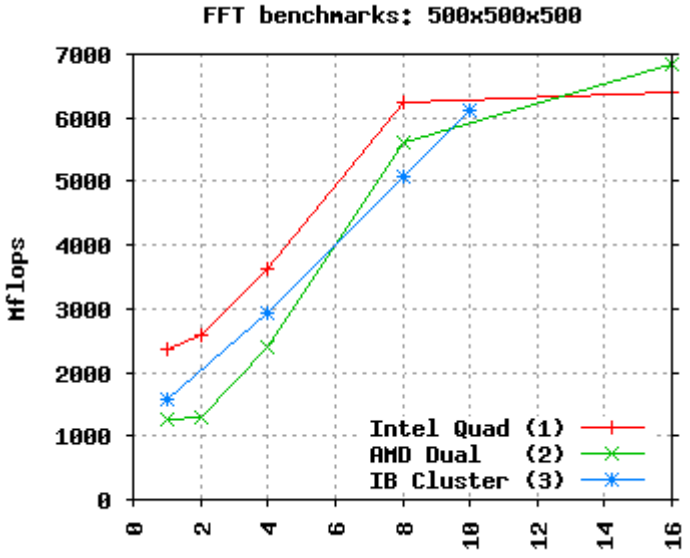
#	CPU	Type	Speed / GHz	OS	kernel	CPUs	Cores / CPU	Nodes	Total Cores	RAM / Core	Interface	Host
1	Intel(R) Xeon(R)	X7350	2.93	SLD5.0	2.6.18	4	4	1	16	4.0 GB	bus	fast06
2	AMD Opteron (tm)	885	2.20	SLD4.4	2.6.9	8	2	1	16	4.0 GB	bus	hasgkss xtmrs
3	AMD Opteron (tm)	252	2.59	SLD5.0	2.6.18	1	2	8	16	1.5 GB	IB	plejade @ifh

- Compute environment:

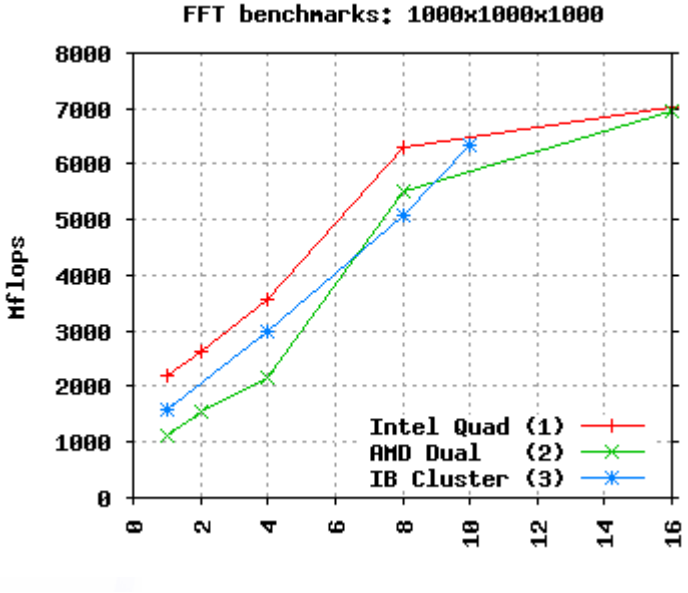
Compiler:	**Intel 10.1, gcc/gnufortran 4.x, PGI 7.1**
MPI:	**openmpi 1.2.5, mpich2 1.0.4**
FFTw:	**fftw 3.2-alpha3**

Comparison





- QuadCore (16 core) ~ 7Gflops → 21s /1000³ FFT
- DualCore (16 core) ~ 7Gflops → 21s /1000³ FFT
- DualCore (IB-Cluster) ~ 7Gflops → 21s /1000³ FFT



Performance summary

HAMBURG • ZEUTHEN

- 16 core DualOpteron offers best scalability & memory throughput

- 16 core QuadCore offers the same performance
- considerably faster for less memory demanding tasks

- Infiniband Performance and Scalability seem sufficient

- Infiniband Cluster of 4*16 sufficient to get execution time << 10s for 1000^3

2. Computing Resources

HAMBURG • ZEUTHEN

- Grid based resources (Linux)
  - very large scale (world-wide distributed)
  - gLite and GT4 Middleware in use
  - Batch like interface (i.e. glite-job-submit), VO driven
  - LHC usage: large scale data reduction, MC, Analysis ...
- Cluster, Farm (Linux) (like NAF or Bird)
  - large scale (site-wide)
  - Batch & Interactive Access (include *remote*)
  - flexible, high availability, easy extend & maintain
  - LHC usage: Physics Analysis, private MC, SW dev

The Grid: potential benefit

HAMBURG • ZEUTHEN

- **Grid-Methods as standard user interface**
- **Resource provider (not free!)**
- **Grid-Services for operations (AAA,…) but security is an issue**
- **Political reasons**

3. Storage Resources

HAMBURG • ZEUTHEN

- Grid based
  - worldwide access (gridftp) / local access (dCap, xrootd)
  - very cost effective large scale *managed storage*
  - optimized for high aggregate tp. with thousands of streams
  - includes *tertiary storage* access/management – large tape systems (robot) these days
- Cluster / Farm
  - cluster filesystem (i.e. Lustre) for very fast and easy access to large datasets, Infiniband based
  - standard file access – like other filesystems
  - access to AFS and Grid storage through IP network
- Workgroup
  - AFS, NFS, CIFS based

Deployed Today

HAMBURG • ZEUTHEN

- **Grid based**
  - **~830 CPU cores (2 GB per core)**
  - **~250 TB disk based (dCache managed) storage**
  - **tape (LTO3/4)**
- **Cluster / Farm**
  - **256 CPU cores (2 GB per core)**
  - **~50 TB Lustre based disk storage**
- **Workgroup**
  - **several TB for homedirs and dedicated group NFS**

todays situation

- **2x SUN SL8500**
- **Installed at DESY in Jan 07**
- **Up to 13000 Cartridges**
- **Multi library capability**
- **Currently 30 drives LTO3, 8 drives LTO4**
- **LTO3 400 GB/Cart, 120 MB/s (45€/cart)**
- **LTO4 800 GB/Cart, 120 MB/s (70€/cart)**
- **2nd Silo financed via grids money**

todays potential

HAMBURG • ZEUTHEN

- **CPU**
  - **Grid managed**
    - **>50000**
  - **Cluster / Farm**
    - **>1000 (exceeding 10,000 soon)**
- **Network**
  - **IP**
    - **WAN – 10Gigabit DFN link in place**
    - **Gigabit, 10Gigabit (and aggregates of such, i.e. 4x10GE)**
  - **Infiniband**
    - **DDR (double data rate) 20 Gbits/sec – per port**
    - **usually 2 ports per node available**

todays potential

HAMBURG • ZEUTHEN

- **Tape**
  - **single stream ~120 MB/sec (native – not compr.)**
  - **aggregate >2 GB/sec**
  - **capacity >8 PB – easy extendable**
- **Disk**
  - **dCache**
    - **unlimited capacity – biggest site >2 PB**
    - **single stream ~100 MB/sec (dep. IP path, disk config)**
    - **aggregate – only limited by network and disk config**
  - **Lustre**
    - **max. observed capacity (worldwide) ~ 2 PB**
    - **bandwidth**
      - **single stream ~1 GB/sec**
      - **aggregate >40 GB/sec (dep. on disk and server config)**

(known) XFEL demands fitting (2013)

HAMBURG • ZEUTHEN

- **DAQ (single Experiment, prob. x 3)**
  - **5 GB/sec aggregate store rate**
    - **todays LTO4 tape drive (extend to ~50 in total)**
    - **extend Ethernet to 10GE per participating node**
    - **need ~50 streams (i.e. 5 per sending node)**
  - **25 PB (50% uptime)**
    - **extend Robot space to ~50000 slots**
      - **5 of the largest robots (reasonable !)**
- **by that time (2013)**
  - **Tape at least: 3.2 TB, 300 MB/sec per drive (x3 better)**
  - **Ethernet (IP): 40 or 100 GE and aggregates of such**

the real big issues...

HAMBURG • ZEUTHEN

- **how/where/when reading data back**
- **long term storage (DPG good practice)**
- **collaborating institutes (copy of data – parts)**
- **analyze use cases (i.e. common data reductions for large groups)**
- **OS platform for analysis (Linux/OSX/Win/...)**

- **this all determine to a very large extend the architecture and costs of the final system**

in other words

HAMBURG • ZEUTHEN

- **need a *Computing Model***
  - **we (IT@DESY) are ready (hot standby) to participate in creating and finalize (iterate)**
  - **must include rough timing (initially)**
  - **must include *Challenges* (prove scaling) for *online* and *offline* processing**
  - **need this also for cost estimates**
    - **somebody will probably ask soon !**

Conclusion

HAMBURG • ZEUTHEN

Still many unknowns

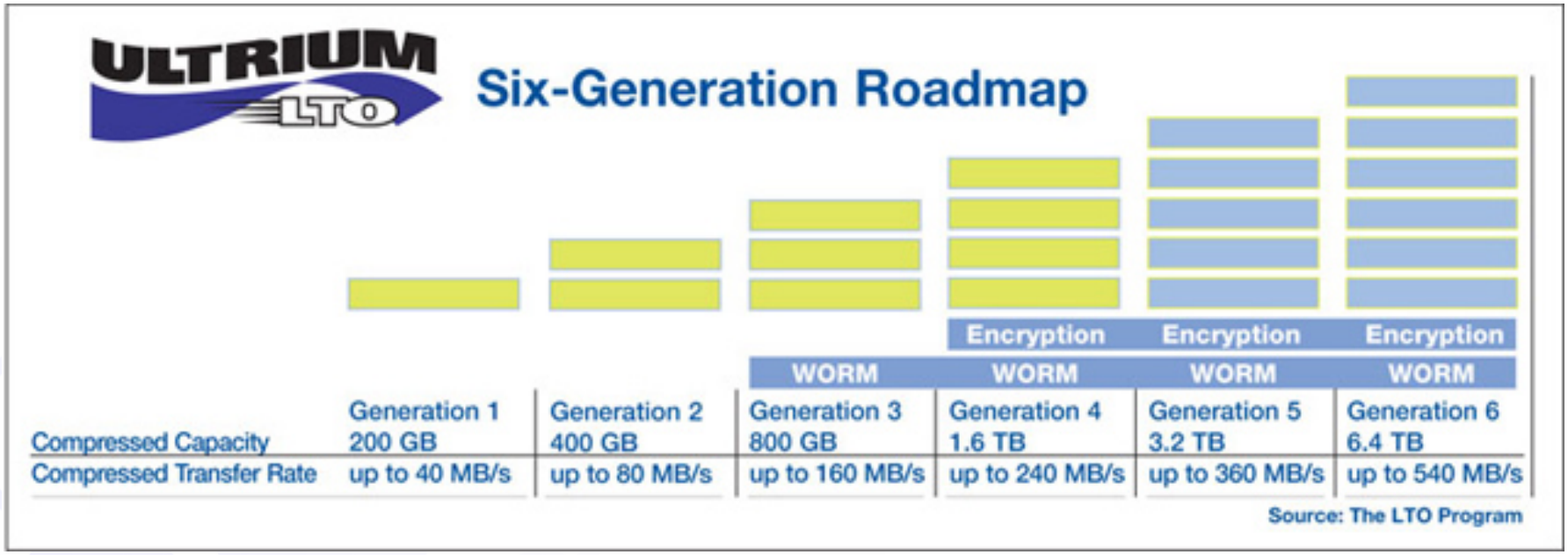
A computing TDR is recommended

It seems as if there are no technical problems

It seems as if it's a matter of money

Software should take multicore architecture into account

technology outlook



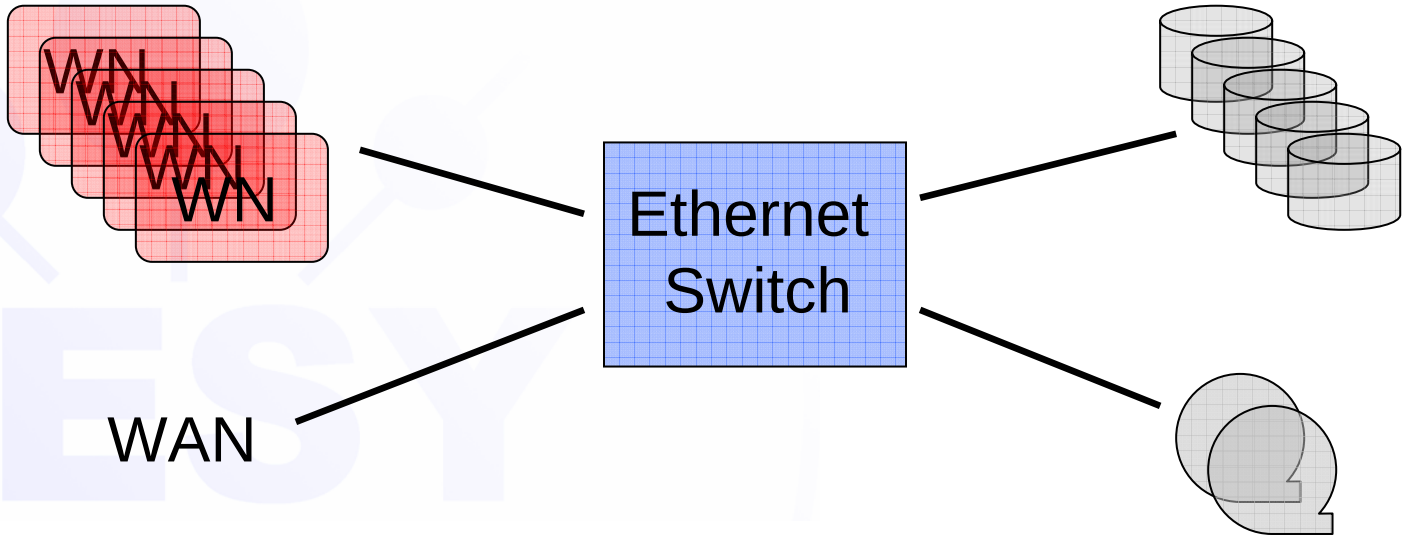
Q4/07 ~2009 ~2012

Note: all numbers include 1:2 compression ratio - divide by two to get real numbers ;-)

simplified architecture Grid fabric

HAMBURG • ZEUTHEN

- **Worker-Node (CPU) + File-Server (Storage) connected through IP Network (GigE) to the (Grid) world (WAN)**
  - **allows commodity components to a large extend**
  - **non commodity: tape, WAN, (LAN)**
- **LCG/gLite Grid middleware stack used**



simplified architecture Cluster/Farm