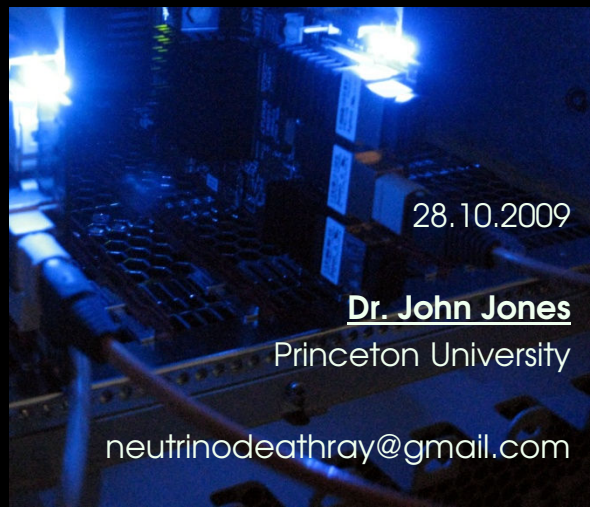
CMS
A Future Trigger Architecture



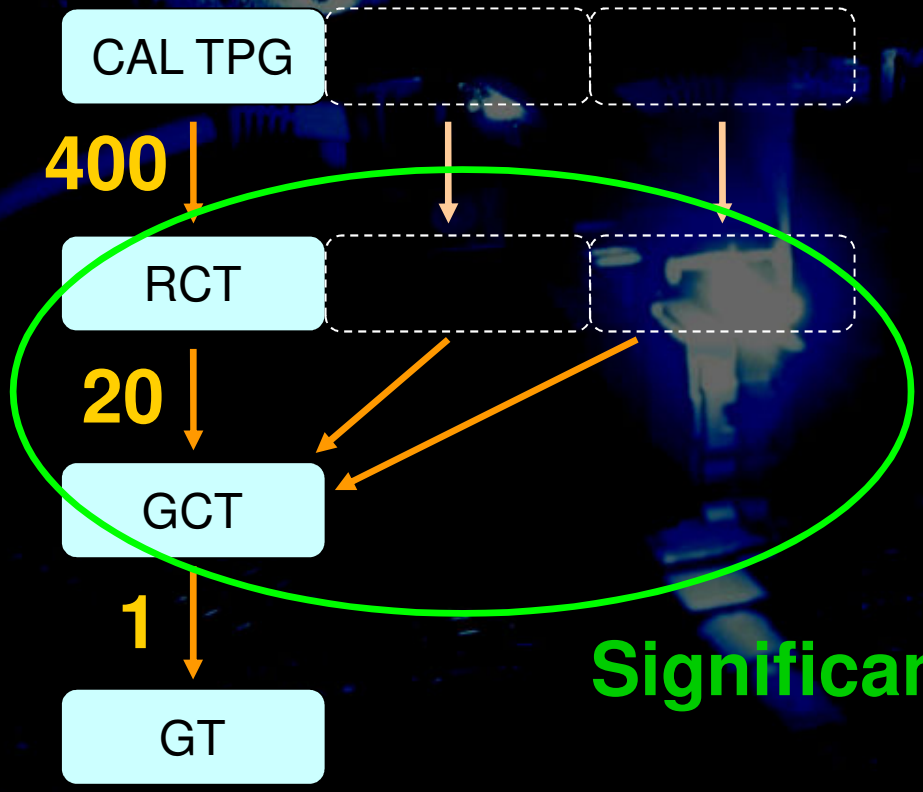
28.10.2009

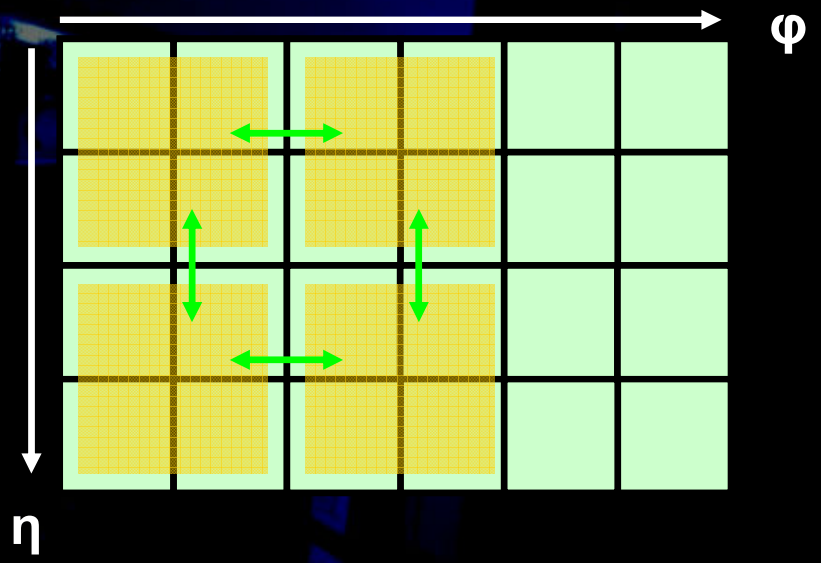
Dr. John Jones
Princeton University

neutrinodeathray@gmail.com

Current CMS Trigger Architecture

Processing subdivided into eta-phi regions / link (e.g. calorimeter trigger)





Significant data reduction

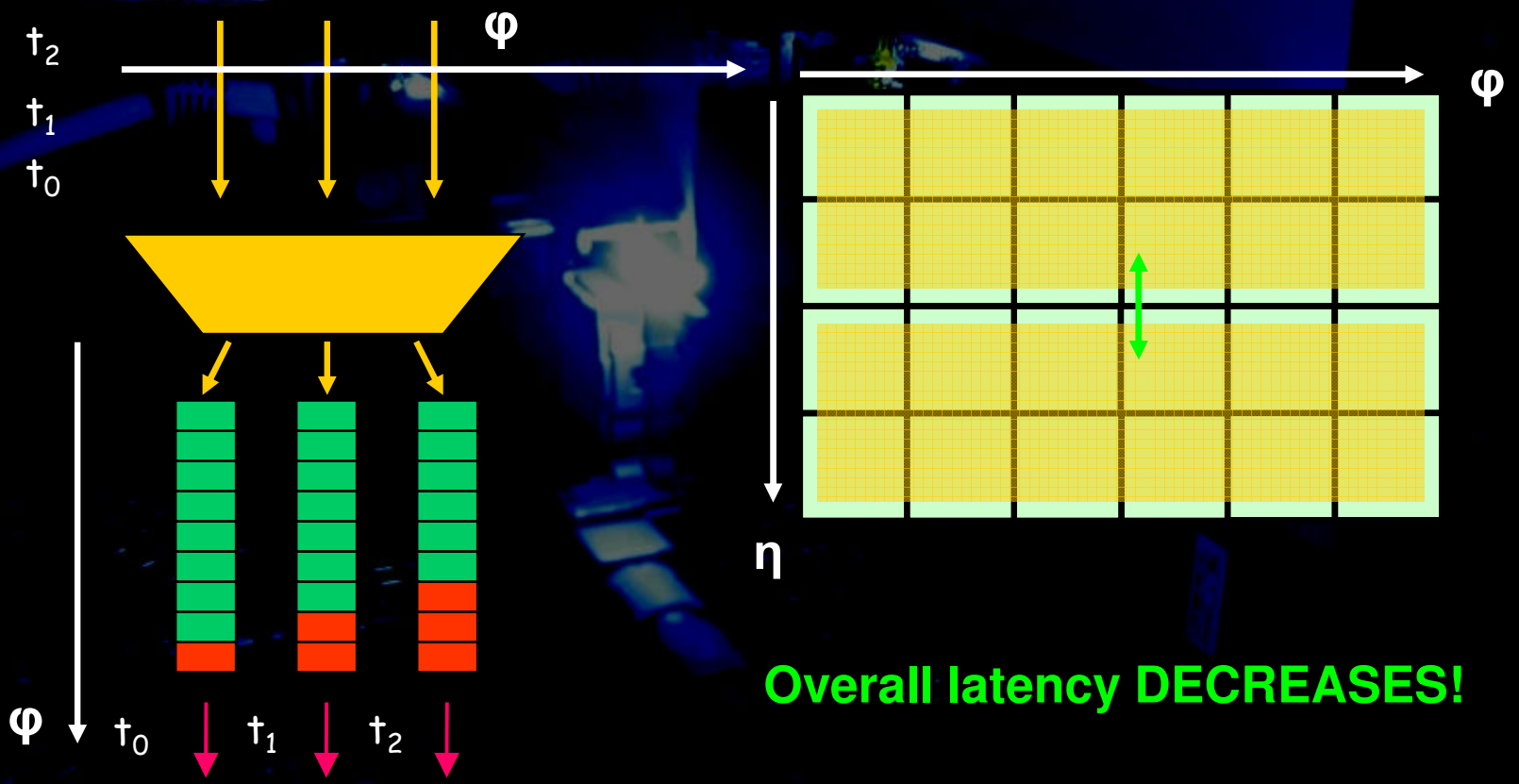
2 scaling problems with this approach:

Difficult to add new input sources (i.e. improved HCAL, tracking)

Data reduction layer doesn't scale efficiently & balance boundary data sharing

Alternative Approach – Time-Multiplexed Data Serialisation

TPG multiplexes data into BX-serialised streams:



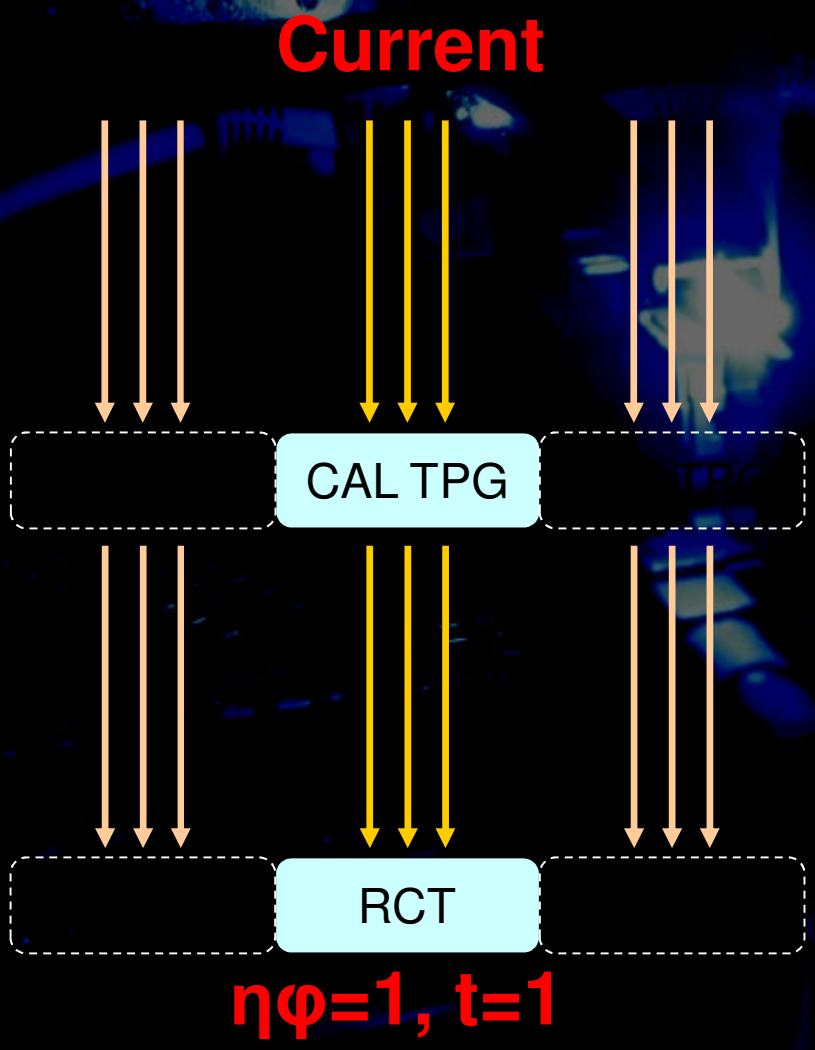
Overall latency DECREASES!

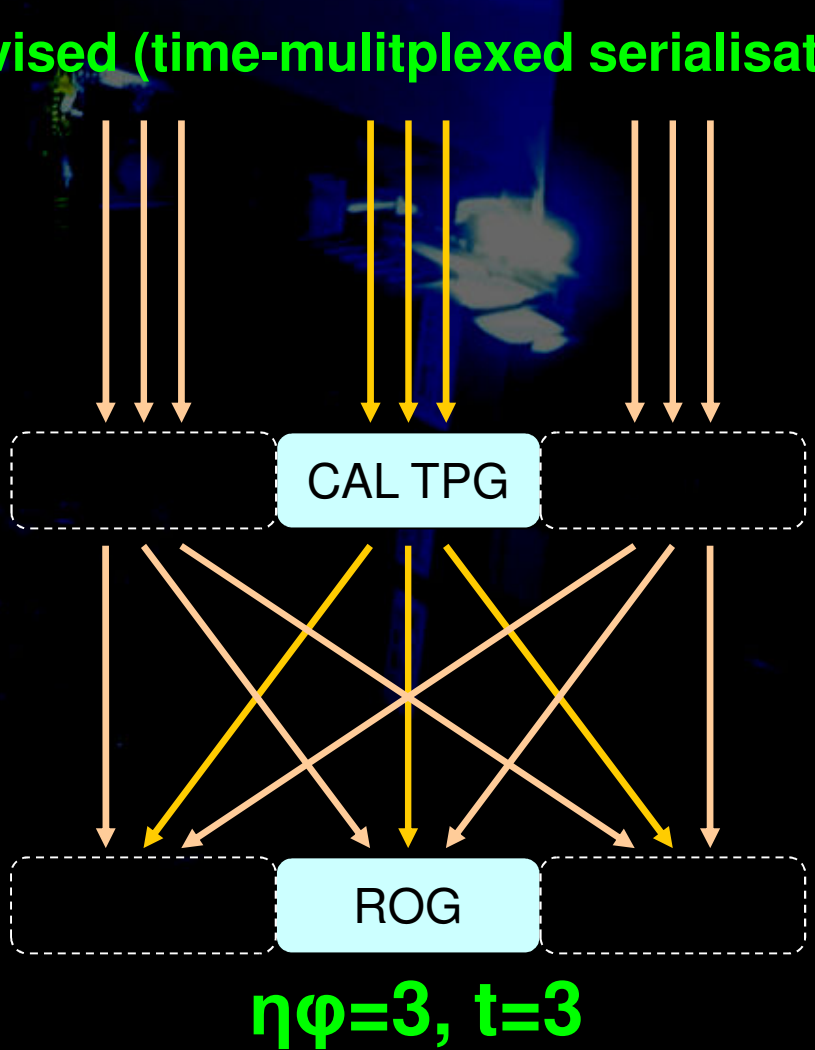
Initial cost: lost time due to multiplexing

Later gain: **Compact, redundant, time-multiplexed system up to GT**

Current vs. Revised Trigger Architecture

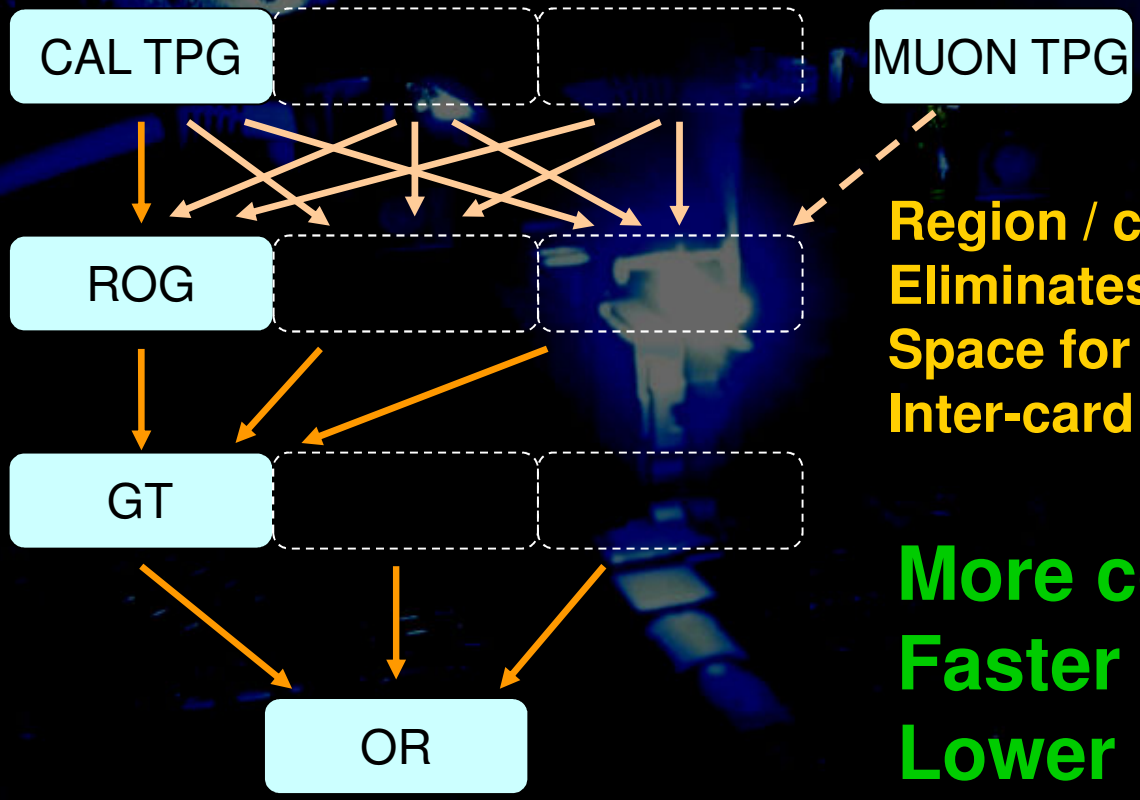
Revisit calorimeter TPG principle:





New Trigger Architecture

Processing subdivided into eta-phi regions / link (e.g. calorimeter trigger)



Region / card increases
Eliminates GCT / RCT boundary
Space for additional future data
Inter-card data sharing decreases

More compact
Faster
Lower latency
Topological
No pre-clustering

Why More Compact?

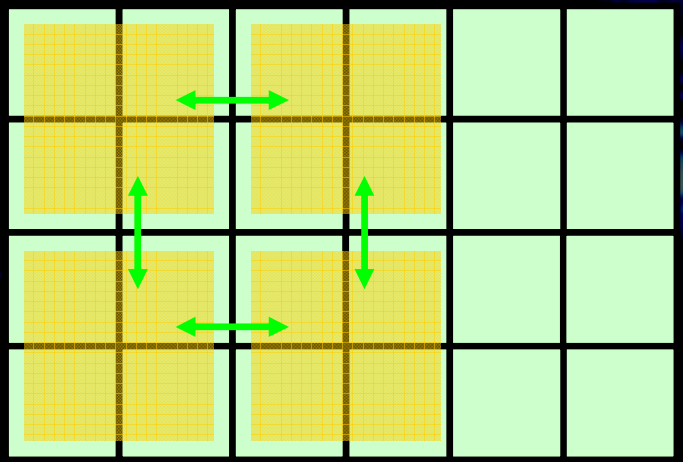
Time-multiplexing folds data-sharing into a single chip

Boundary sharing becomes a pipelining issue

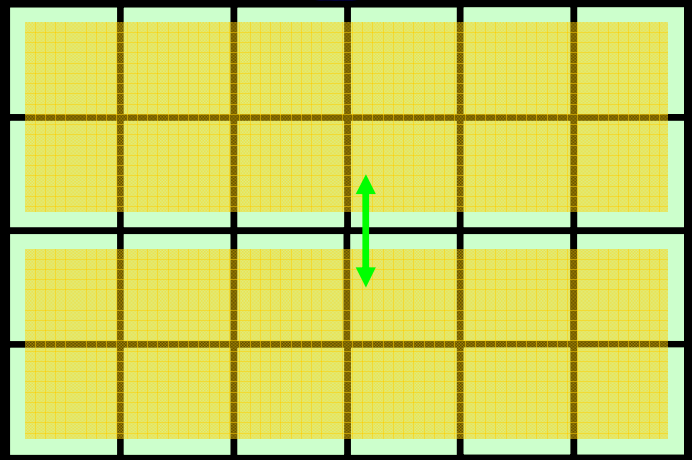
Boundary sharing creates a geometric, highly non-linear increase in resource usage

Reducing sharing even a small amount has a significant effect

You can avoid crossing between crates…



c.f.



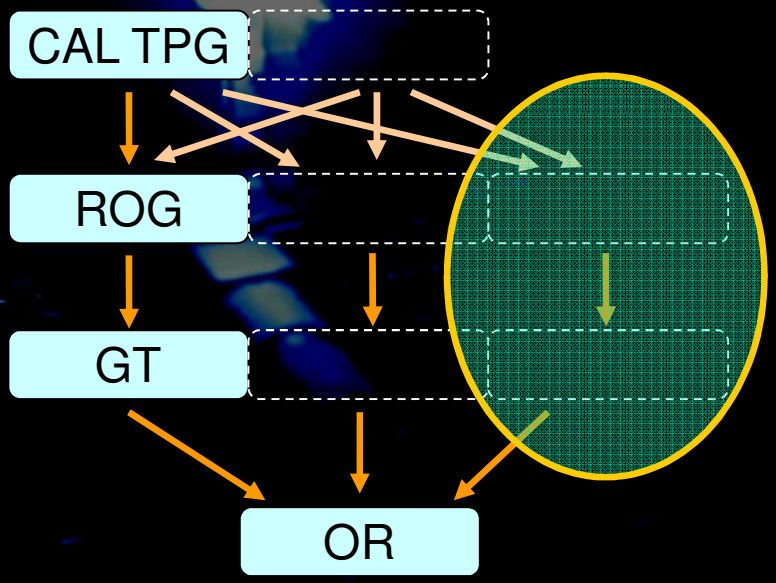
Why Redundant?

Time multiplexing makes it very easy to provide data duplicates

A crate failure causes dead-time, not regional physics loss

A spare can be incorporated using extra fibers from the TPGs

Can be switched over during a run…



Spare processing crate

Why Faster / Lower Latency?

Eliminating data sharing condenses result sorting

- Topological processing can occur at an earlier stage

Again, this is a geometric problem

- The benefit is far better than linear
- Synthesis tools 'cope' far better with pipelined designs

Initial pipelining latency improves future processing performance

Doing the Numbers (Based on Current CT)

Post-TPG link speed ~3.75Gb/s ~8b * 9.375 / BX / fibre
16 x serialisation in TPG => ~75 towers (ECAL+HCAL) / BX / fibre

Eliminate phi-boundary (*one fibre absorbs entire eta segment!*)

Calorimeter dimensions 88 (eta) x 72 (phi) trigger towers
e.g. 1 matrix card = 16 (eta) x 72 (phi)

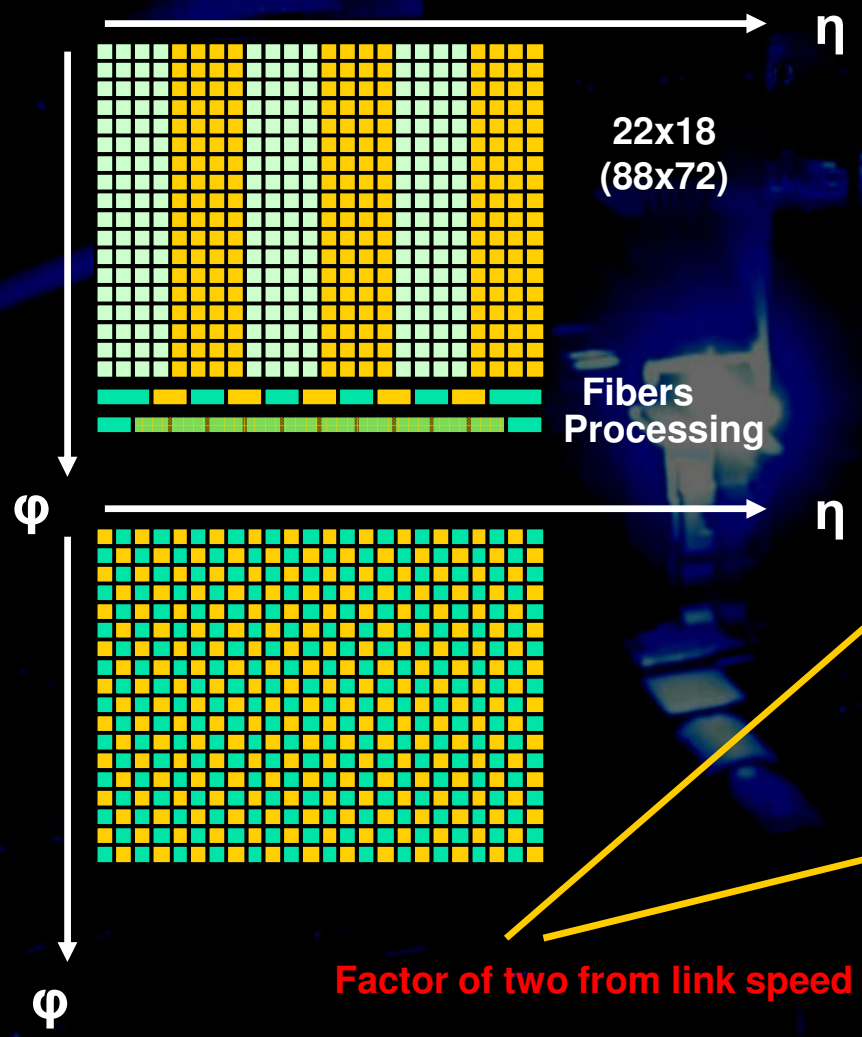
16 input channels => all inputs for jet trigger + overlap in current CMS

10 matrix cards for full-phi-granularity, coarse (4 tower) eta processing (x16 copies)
16 matrix cards for full-tower-granularity processing (x16 copies)

2 fibers => output for results (electrons, photons & jets)

32 input fibres into GT card

Processing Topology – New and Old



New Scheme
3x3 jet tower finder (full phi resolution)
4x4 calorimeter towers / jet tower
3.75Gb/s links

Data sharing – input fiber ratio: 160/88 = 1.82

Real input fiber count: 16x88 = ~1408

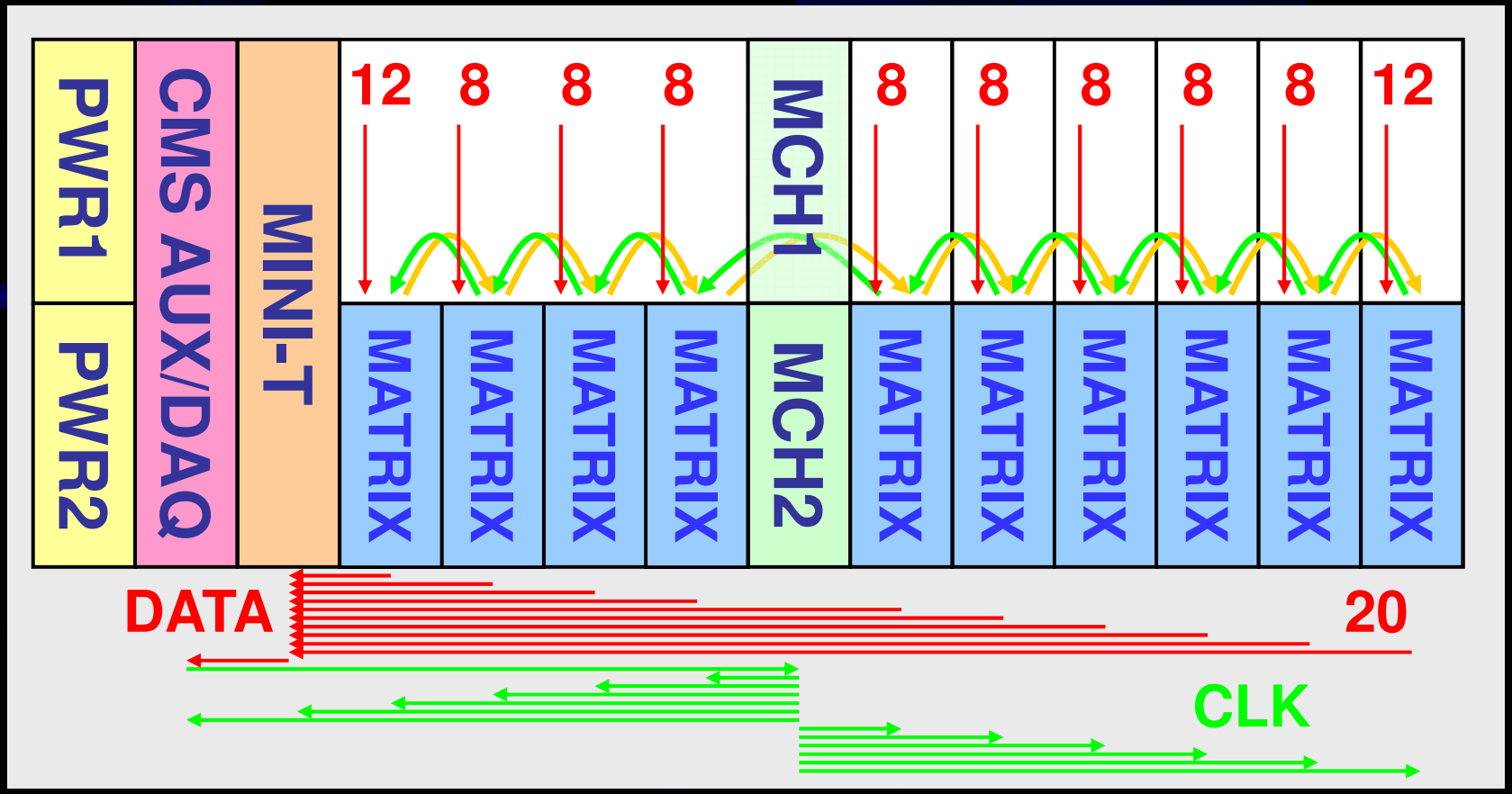
Old scheme – NN sharing
6.5Gb/s links

Data sharing – input fiber ratio: ~21888/680 = 32.19

Real input fiber count: 16x72x88/144 = ~680

Factor of two from link speed – need 6.5Gb/s to use old scheme

The Modular Trigger Crate – 3.75Gb/s, Partial Granularity



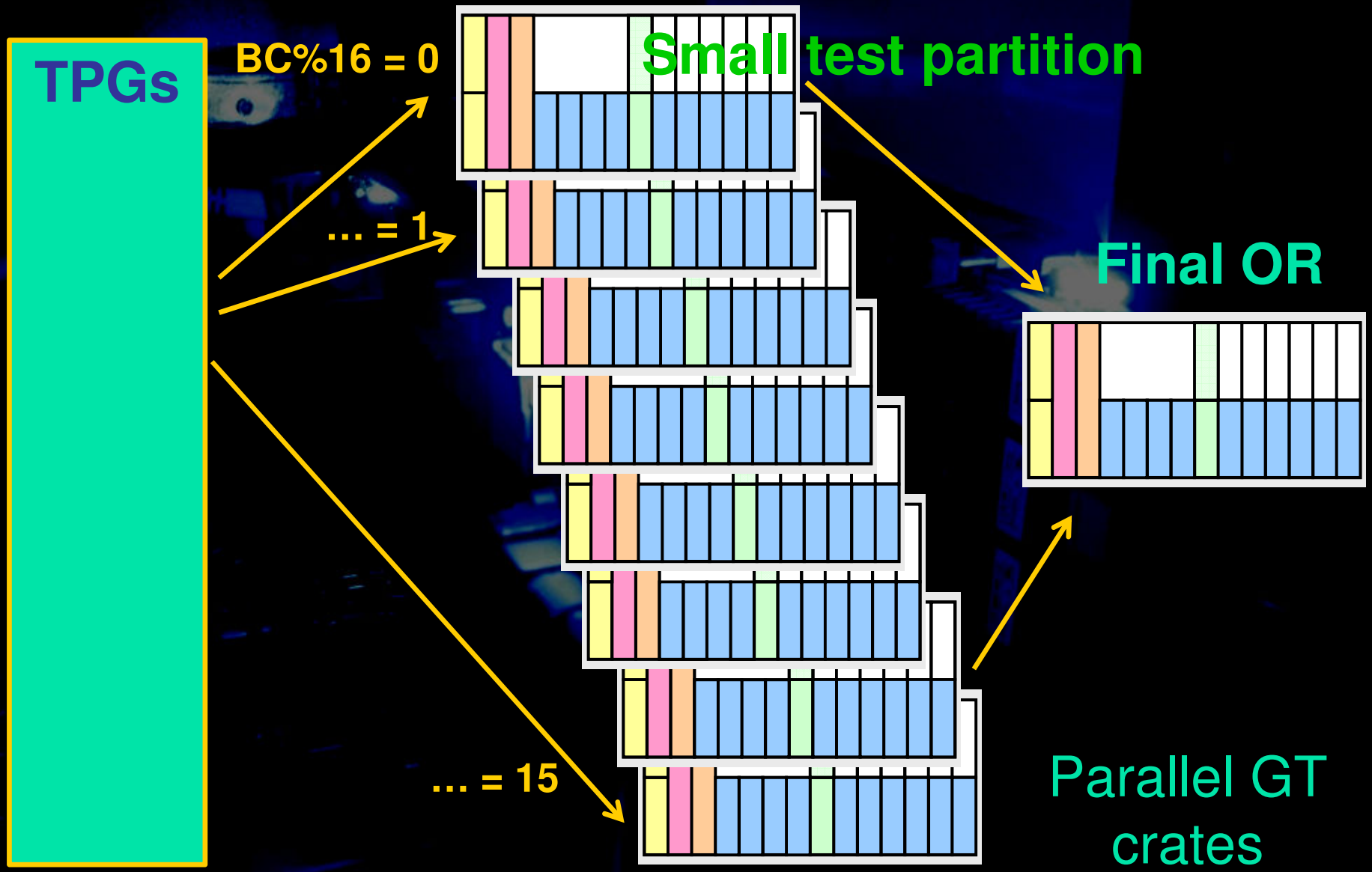
Can have a fully-redundant crate (spare fibres from TPG)

Redundant power & communications

Improvements in link speed = reduction in crate size *or* latency

Complete system test can be achieved with a small setup (e.g. debug)

Multiplexed Framework



An Example, The Current GCT

Currently algorithms include:

- Jets
- Taus
- MET
- MHT
- Sum ET
- Sum HT

A new processing implementation using time-multiplexed data flow

- Parallel in eta, serial in phi

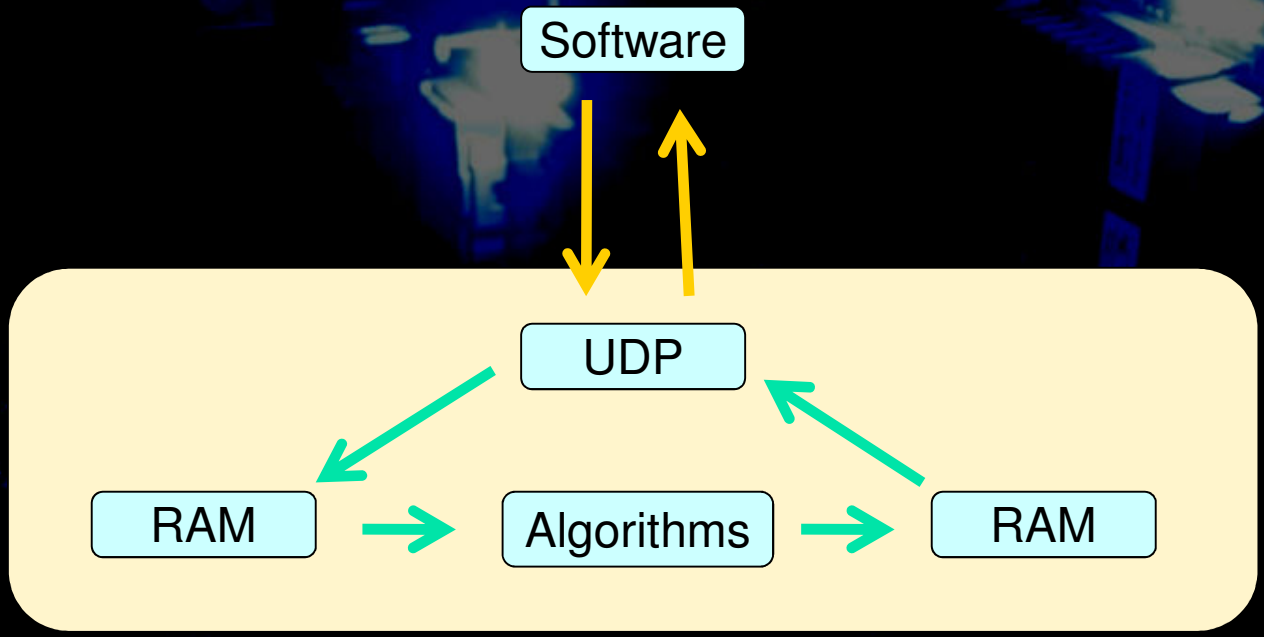
As the data is serialised in phi, so processing algorithms should be pipelined

- Each algorithm can be split into parts:
- Jet sum
- Tau veto bits
- Jet maximum

FLI-UDP Bridge

Software-firmware test bench

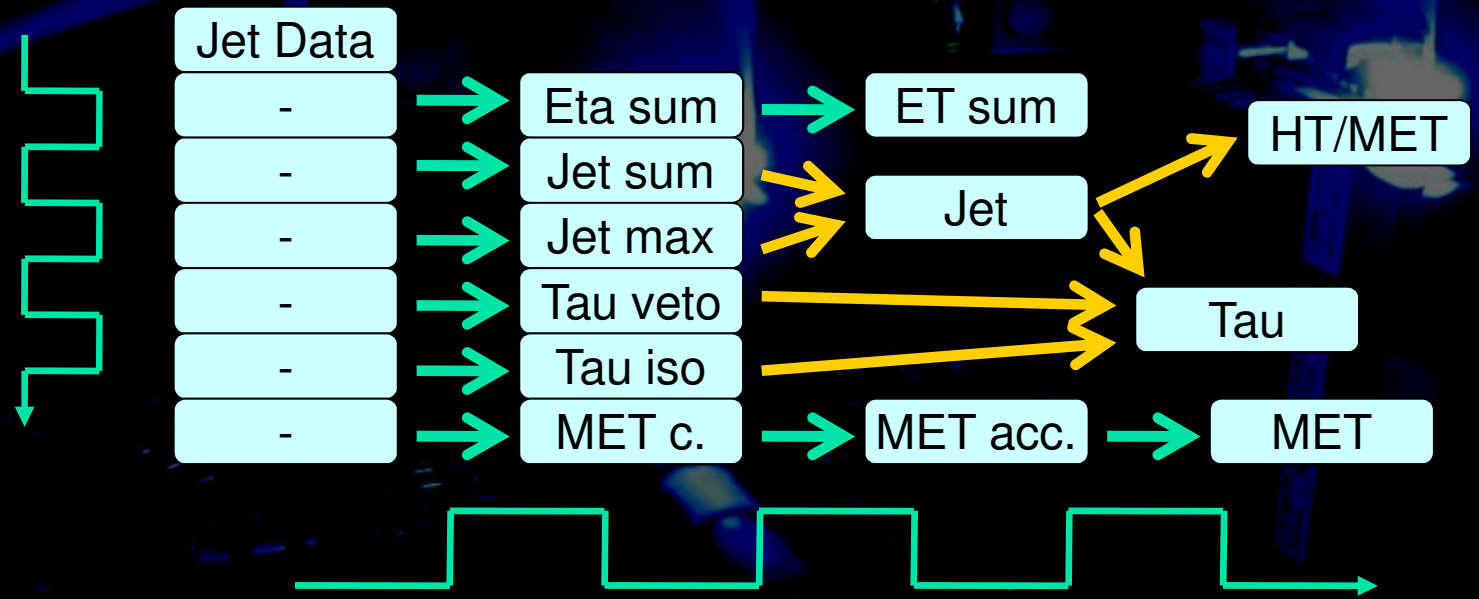
- Allows direct testing of algorithms using 'virtual' FPGA
- Software control layer is identical to the one in the final system



Pipelined Architecture

Split processing into many micro-operations

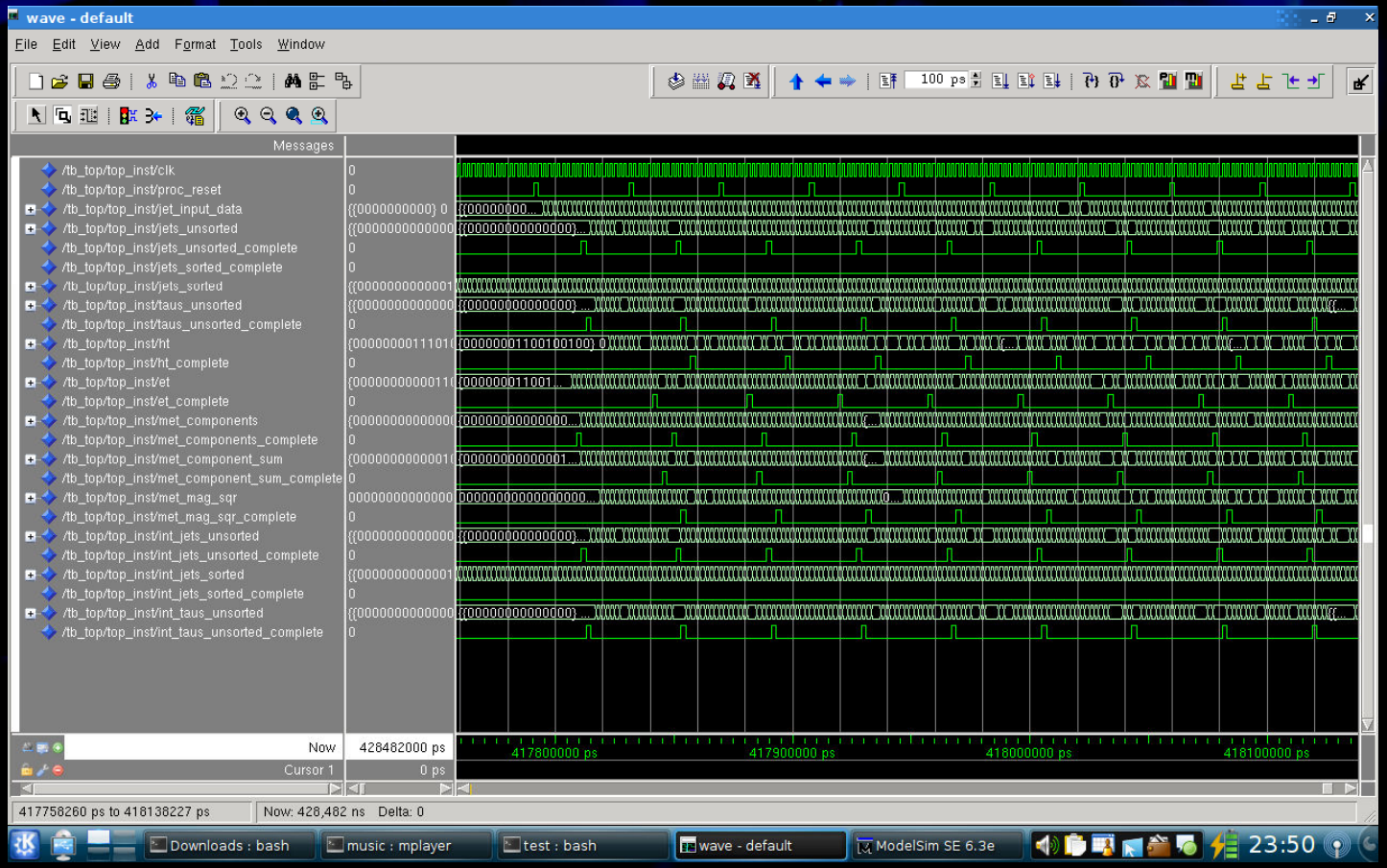
Each micro-operation will easily meet timing at > 300MHz



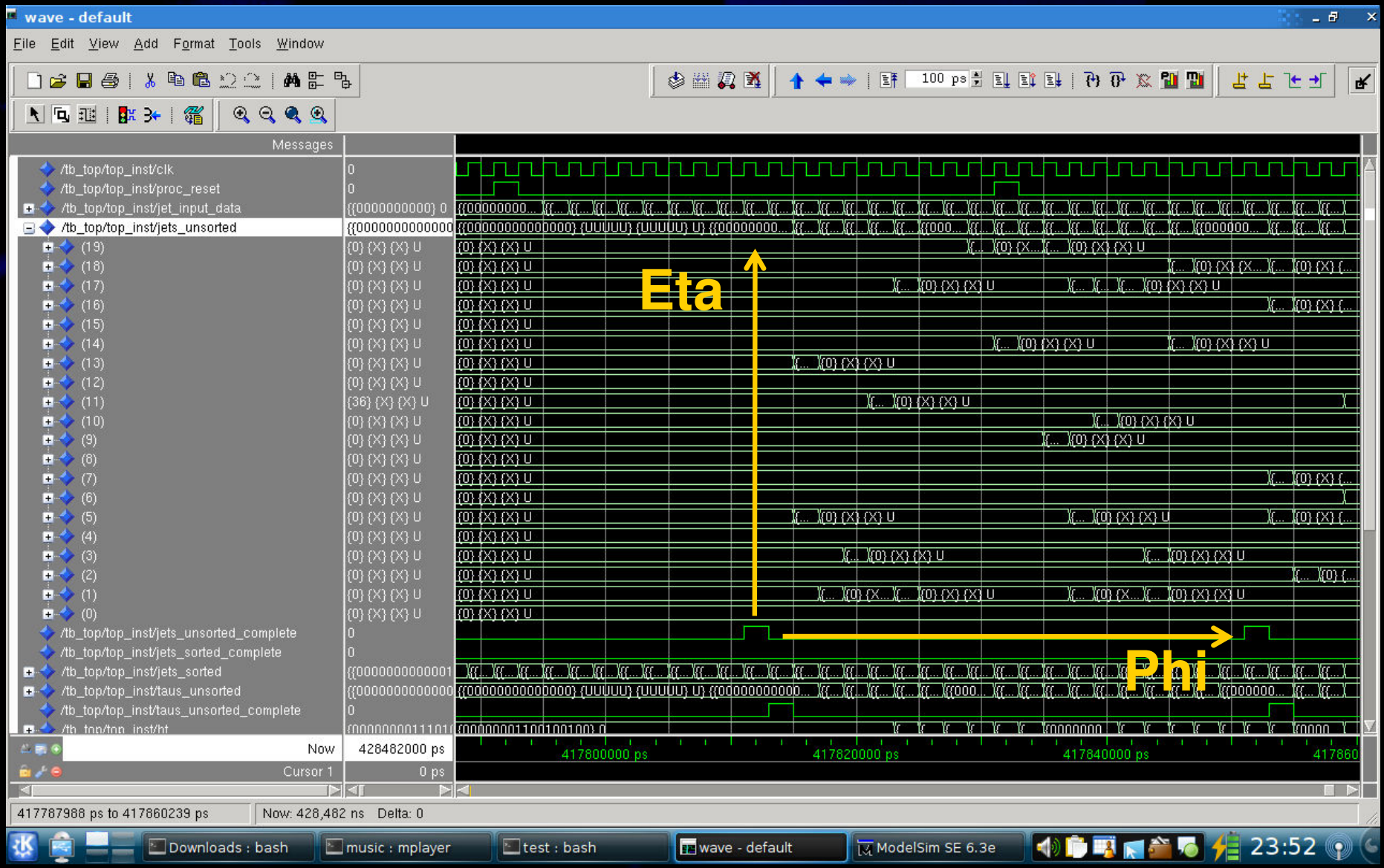
Results I

All components implemented and synthesised in the Matrix card

- Maximum clock speed ~300MHz (could be optimised further)
- The entire GCT fits in 8% of a single matrix card
- Latency ~5BX



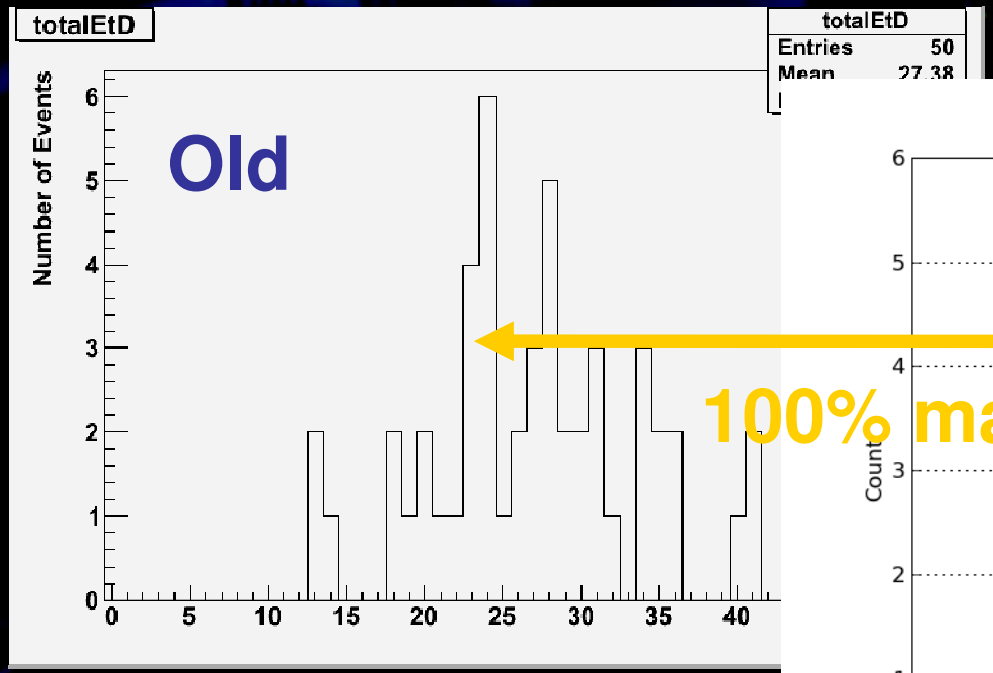
Results II

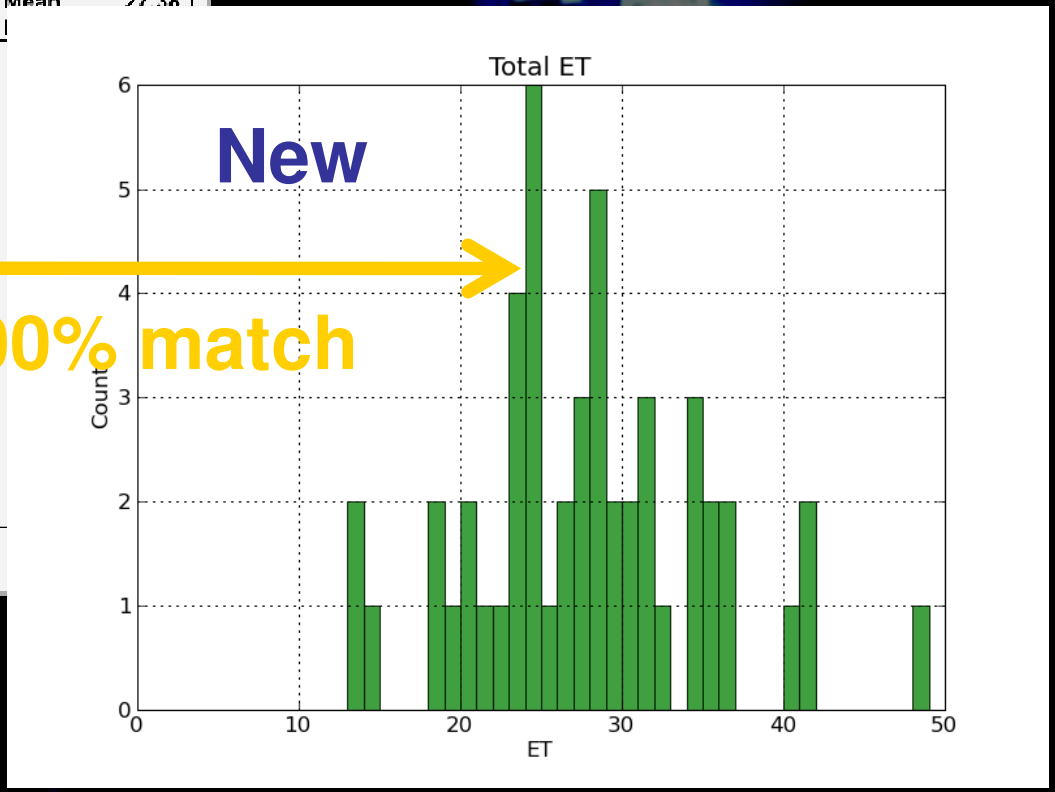

wave - default
File Edit View Add Format Tools Window
Messages
/tb_top/top_inst/clk
/tb_top/top_inst/proc_reset
/tb_top/top_inst/jet_input_data
/tb_top/top_inst/jets_unsorted
/tb_top/top_inst/jets_unsorted_complete
/tb_top/top_inst/jets_sorted_complete
/tb_top/top_inst/jets_sorted
/tb_top/top_inst/taus_unsorted
/tb_top/top_inst/taus_unsorted_complete
Eta
Phi
Now
428482000 ps
Cursor 1
0 ps
417800000 ps
417820000 ps
417840000 ps
417860
417787988 ps to 417860239 ps
Now: 428,482 ns Delta: 0
Downloads : bash
music : mplayer
test : bash
wave - default
ModelSim SE 6.3e
23:52

Results III

Compare current GCT emulator with new algorithms

- Results are identical (except for binning, etc…) in ET sum case
- Data shown in slide from CRAFT





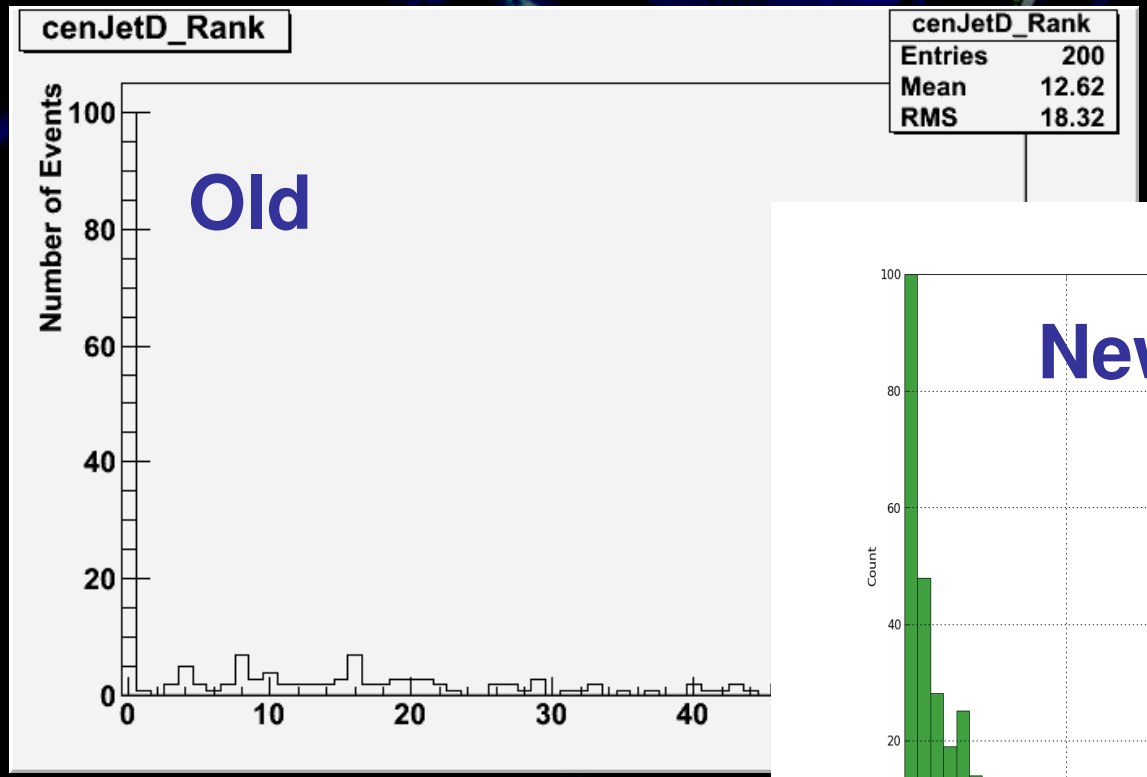
100% match

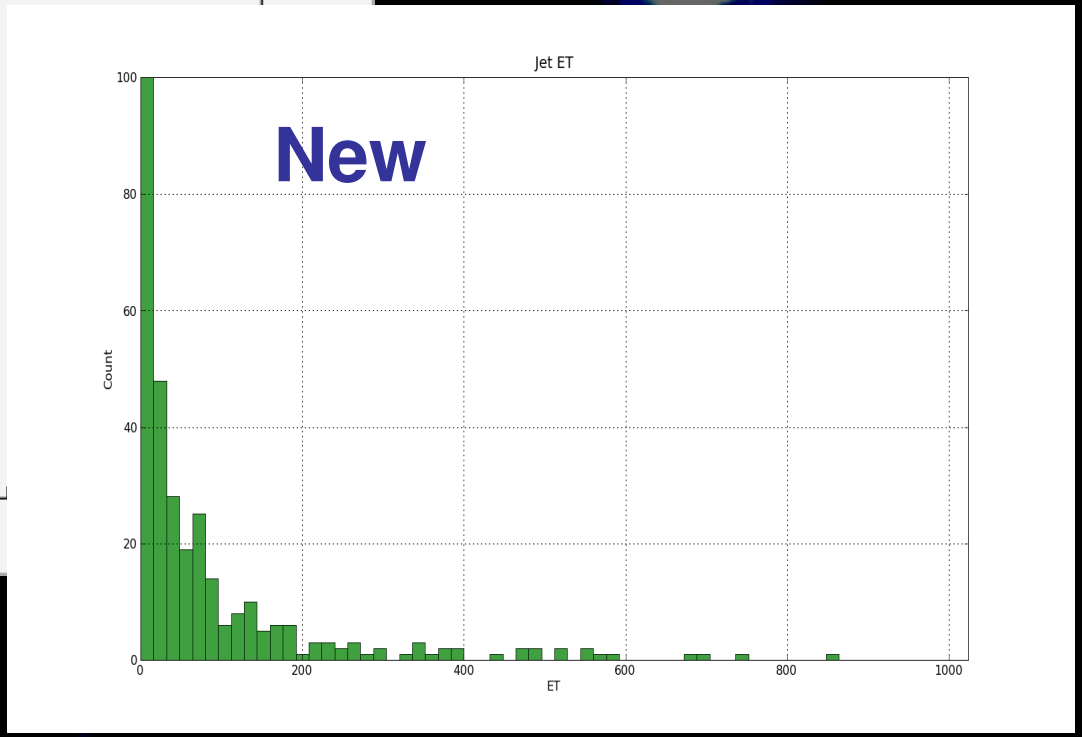
Results IV

Some minor differences to current algorithms for Jet Et (SUSY LM0 dataset)

Jet calculations are performed in a different way

New version shows all jets, no pre-clustering, no rank conversion



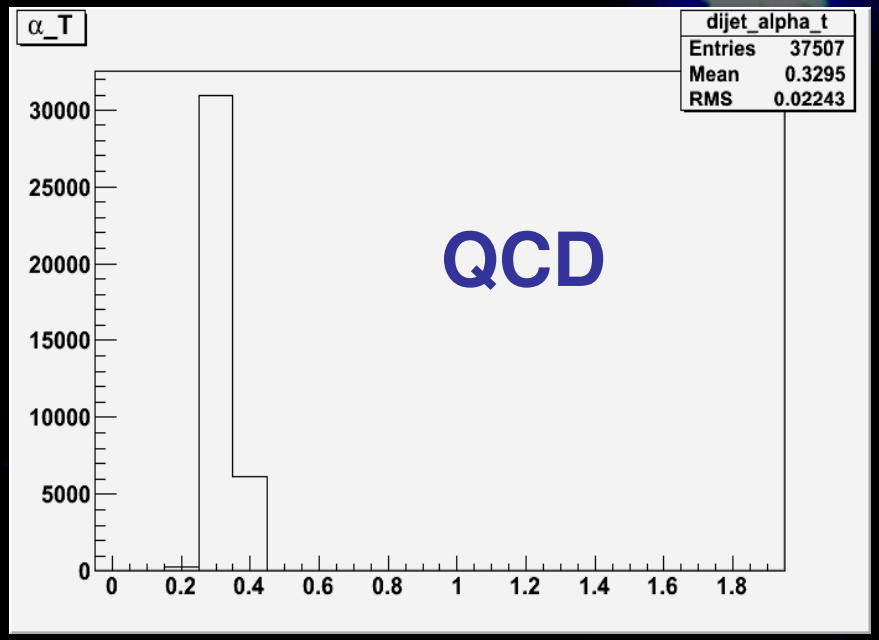


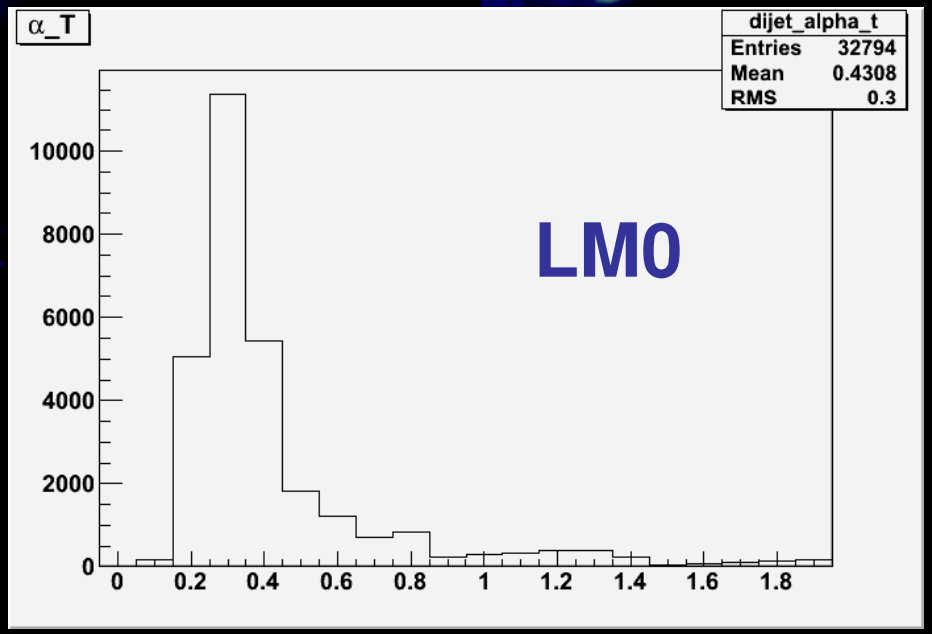
Results V – Alpha T Squared

Going beyond current algorithms, highly topological example

- Used in CMS SUSY analyses (https://twiki.cern.ch/twiki/bin/view/CMS/SusyDijet)
- Useful as soon as jets <= 50GeV are prescaled
- Implemented by Jad Marrouche in VHDL

Preliminary result (note values are squared):





The M^2 (Matrix Squared)

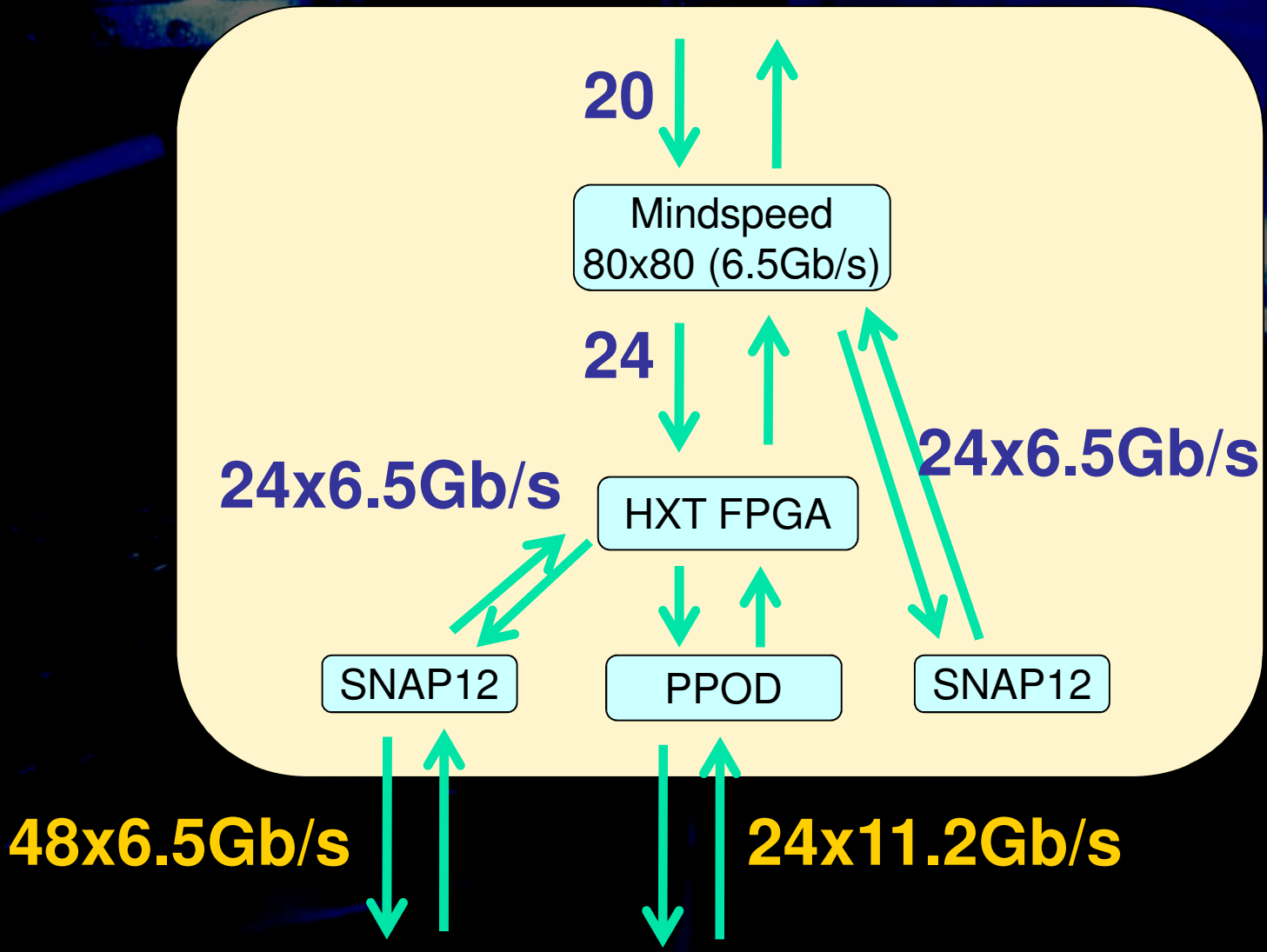
Successor to the Mini-T and the Matrix card
- Converge the two designs and push to the limit of current technology
- Logical place to build the new trigger system (eventually…)

Some similarities with both designs
- 6.5Gb/s cross-point switch
- 11.2Gb/s direct links (PPOD-FP24GA)
- More advanced clock recovery (DPLL – can shift reference to match source)
- DDR2 SDRAM
- Mix of PPOD and SNAP12 optics
- Xilinx Virtex-6 HXT FPGA
- Total bandwidth 11.2x24 + 6.5x48 => ~580Gb/s

Expected to manufacture at the end of 2010 at the earliest

Wait for technology to evolve…

The M² (Matrix Squared)



Conclusions

Time-multiplexed processing provides many advantages:

- Avoids pre-clustering
- Lower latency
- Lower resource usage
- Higher clock speed
- Redundancy
- Small test system

Demonstration implementation shows that the entire current GCT will fit in one matrix card this way...

… if only we had more serial links…

M^2 is expected to be our final implementation platform for first upgrades