INFIERI Summer School
USP, São Paulo, Brazil
23 Jan - 5 Feb 2017

See D. Cieri and L. Calligaris Posters!

Scalability of the TMTT CMS L1 Track Trigger System

C. Amstutz[1], L. E. Ardila Perez[1,4,5], F. A. Ball[2], M. N. Balzer[1], J. Brooke[2], M. Caselle[2], L. Calligaris[4,5], D. Cieri[2,4,5], E. J. Clement[1], G. Hall[3], T. R. Harbaum[1], K. Harder[4], P. R. Hobson[6], G. M. Iles[3], T. James[3], K. Manolopoulos[4], T. Matsushita[7], A. D. Morton[6], D. Newbold[2,4], S. Paramesvaran[2], M. Pesaresi[3], I. D. Reid[6], A. W. Rose[3], O. Sander[1], T. Schuh[1], C. Shepherd-Themistocleous[4], A. Shtipliyski[3], S. P. Summers[3], A. Tapper[3], I. Tomalin[4], K. Uchida[3], P. Vichoudis[8], M. Weber[1] **for the CMS collaborators**

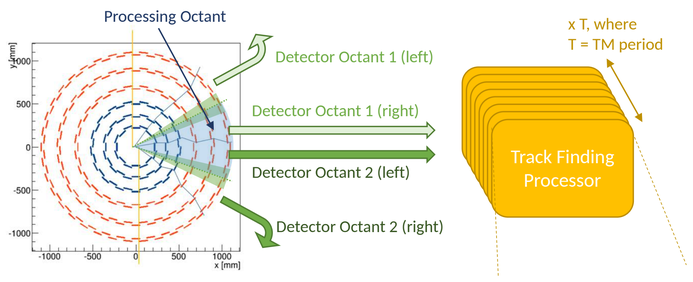
[1]Karlsruhe Institute of Technology,Karlsruhe, Germany [3]Imperial College London, London,UK [6]Brunel University, London, Uxbridge, UK [8]CERN, Geneva, Switzerland
[2]University of Bristol, Bristol, UK [4]Rutherford Appleton Laboratory, Didcot, UK [7]Österreichische Akademie der Wissenschaften, Vienna, Austria
[5]Supported by the EU FP7-PEOPLE-2012-ITN project nr 317446, INFIERI, "Intelligent Fast Interconnected and Efficient Devices for Frontier Exploitation in Research and Industry

Demonstrated scalable slice

1 slice = 1 Track Finding Processor

each slice processes 1/8 tracker in φ

and 1/T in time, where T = Time Multiplexing period



maximum T then determined by $N_{DTC\_links\_out}/2$

- DTC output bandwidth fixed at 600Gbps

link speed	$N_{DTC\_links\_out}$	T
10Gbps	72	36 BX
16Gbps	36	18 BX
25Gbps	24	12 BX

to process 1/8 of tracker in φ, scalable slice receives 1 link/DTC from all DTCs in 2 adjacent octants

- 2 DTC links go to each TFP (1/octant)

Each box is one MP7

Demonstrated scalable slice interfaces

from DTCs

- number of input links is 2 x $N_{DTCs\_per\_octant}$
- geometry dependent (more modules in TP geom = more DTCs)
- up to 105 48b input stubs per link, fixed since TM period proportional to (1/link speed)

72 input links for TP geometry (36 DTCs per octant)

or

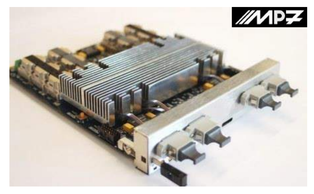
64 input links for Tilted geometry (32 DTCs per octant)

to L1 correlator

- depends on line rate/encoding out
- up to 100 128b tracks per octant/event
- can tune #links vs latency

8 links @ 10Gbps, 8b10b, or
4 links @ 16Gbps, 64b66b, or
3 links? @ 25Gbps, 64b66b

Demonstrator Hardware - the master processor 7 (mp7)



One TFP slice consist of 5 MP7s, a generic high-performance µTCA card developed for the CMS Phase I calorimeter trigger upgrade

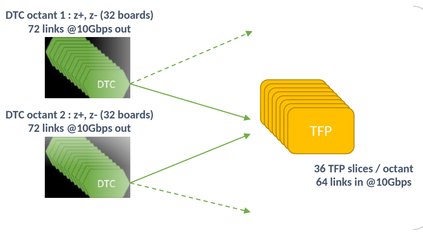
Equipped with a Xilinx Virtex-7 690 FPGA

Has 12 Avago Technologies MiniPOD optical transceivers each providing 12 channels at 10.3 Gbps each

Full system - conservative scenario

what would system look like if we took the current demonstrator and scaled it up as is

- assumes **no change** in algorithms/performance necessary & no optimisation of design
- design is untouched, **no 'f/w scaling' necessary**
- assumes use of **10Gbps links** only
- assumes TFP boards would use currently available FPGAs, and reasonably conservative design
- low-risk but obviously over-conservative, given the timescales

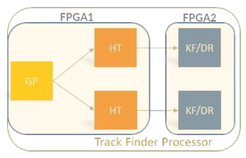


link speed	10 Gbps
T	36 BX
$N_{DTC\_links\_out}$	72
$N_{TFP\_links\_in}$	64
$N_{TFP\_links\_out}$	8
N_{TFPs}	288 (8x36)

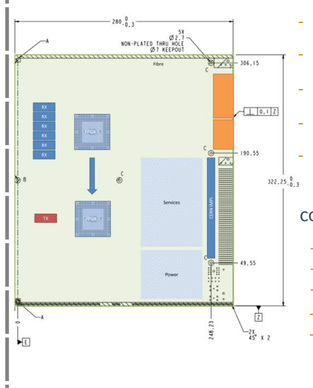
Board design

take demonstrator and split into two

- each part of the design fits reasonably into a **Kintex Ultra115**
- 60-80% utilisation (inc. infra) assuming no changes to design
- 64 input links to FPGA1, 8 output links from FPGA2
- unidirectional *high speed serial bus* between FPGA1 and FPGA2



	LUTs	FFs	BRAM36	DSP48	
GP	128k	228k	198	1560	FPGA1
HT	284k	378k	1368	0	
KF/DR	382k	794k	1750	5040	FPGA2



- optical bandwidth in is **~640Gbps**; out < 80Gbps
- no specific use of backplane
- power per board estimated **< 150W**
- **12 boards/shelf**
- **24 shelves or 12 racks**

conclusion

- low power, low I/O density & not top-of-the-range FPGA
- optical bandwidth approximately that of DTC
- margin in power, board real estate and FPGA logic
- PCB/assembly costs mainly in NRE, less important for production
- infrastructure costs are non-negligible (~1.5 MCHF)

System Cost

scaling up to full system, and including:

- 85% yield on board production
- 10% spare boards and spare ATCA shelves
- ATCA backplanes and enclosures (plus spare)
- hub cards/Blade13 (estimated)
- patch panels, bulk power (AC->DC48V), fibres, rack PCs

	cost (CHF)
FPGA	7.3k
optics	4.8k
PCB/assembly/ components	2.6k
cost/board	**14.7k**

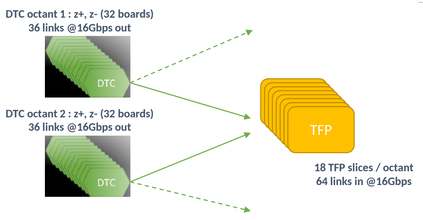
production cost => ~6.9 MCHF

Full system - reasonable/baseline scenario

what would system based on a scaled demonstrator look like in 2025 if...

- we assume use of **16Gbps links**
- demonstrator design is updated for scaled architecture
- assumes TFP boards would use currently available FPGAs, but more/less conservative design (if needed)

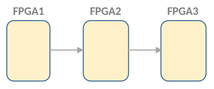
=> effectively asks '*what additional design work is needed in order to scale with system?*'



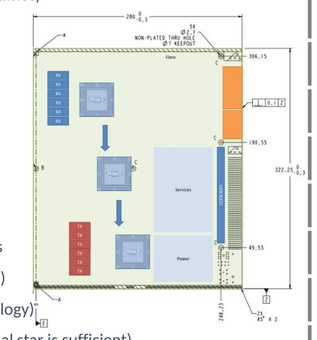
link speed	16 Gbps
T	18 BX
$N_{DTC\_links\_out}$	36
$N_{TFP\_links\_in}$	64
$N_{TFP\_links\_out}$	4
N_{TFPs}	144 (8x18)

Board design

- how to ensure algorithms/firmware compatible with 18BX architecture?
  - same volume of data at twice the rate => **needs 2x processing capacity**
- sum total of logic needed would be **O(3-4) KU115 FPGAs**
  - 60-80% utilisation with no change to algorithms
  - includes estimated infrastructure
  - BRAM reductions likely anyway (currently buffering multiple times)



	LUTs	FFs	BRAM36	DSP48
GP/HT/ KF/DR x2 + infra	1690k	2900k	6830	13200
# KU115	2.5	2.2	3.2	2.3



baseline scenario with 3 FPGAs

- optical bandwidth in is **~1024Gbps** (similar to MP7); out < 80Gbps
- 64 input links to FPGA1, 8 output links from FPGA2/3 (D1517 part)
- unidirectional high speed serial bus between FPGAs (flexible topology)
- no specific use of backplane or backplane bandwidth (e.g. 40G dual star is sufficient)
  - except maybe links to DAQ?
- power per board estimated **<200 W**
- **9 boards/shelf**
- **16 shelves or 8 racks**
- preserve **successful features of demonstrator** (flexibility, scaling, daisy chained 'blocks')

conclusion – a more compact system and a flexible board, well adapted to production timescale

- still low power, low I/O density & not top-of-the-range FPGA
- margin in power (inc. per shelf)

System Cost

scaling up to full system, and including:

- 85% yield on board production
- 10% spare boards and spare ATCA shelves
- ATCA backplanes and enclosures (plus spare)
- hub cards/Blade13 (estimated)
- patch panels, bulk power (AC->DC48V), fibres, rack PCs

	cost (CHF)
FPGA	10.9k
optics	4.8k
PCB/assembly/ components	2.6k
cost/board	**18.3k**

production cost => ~4.3 MCHF

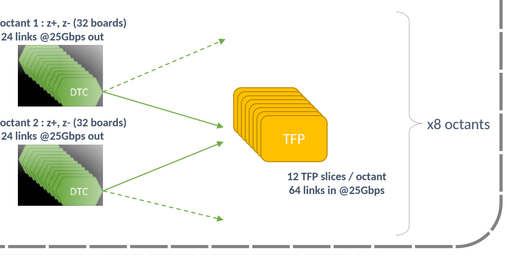
Full system - advanced scenario

what would system based on scaled demonstrator look like if we assume the use of 25Gbps links

what would a 25G system imply?

- smaller system, cost overheads smaller, faster production
- 25Gbps optics is new technology, no experience yet, prototyping required
- not cost optimal? only a large, mid range Virtex Ultracale+, -2, provides sufficient logic and links/link speed

link speed	25 Gbps
T	12 BX
$N_{DTC\_links\_out}$	24
$N_{TFP\_links\_in}$	64
$N_{TFP\_links\_out}$	2-3
N_{TFPs}	96 (8x12)



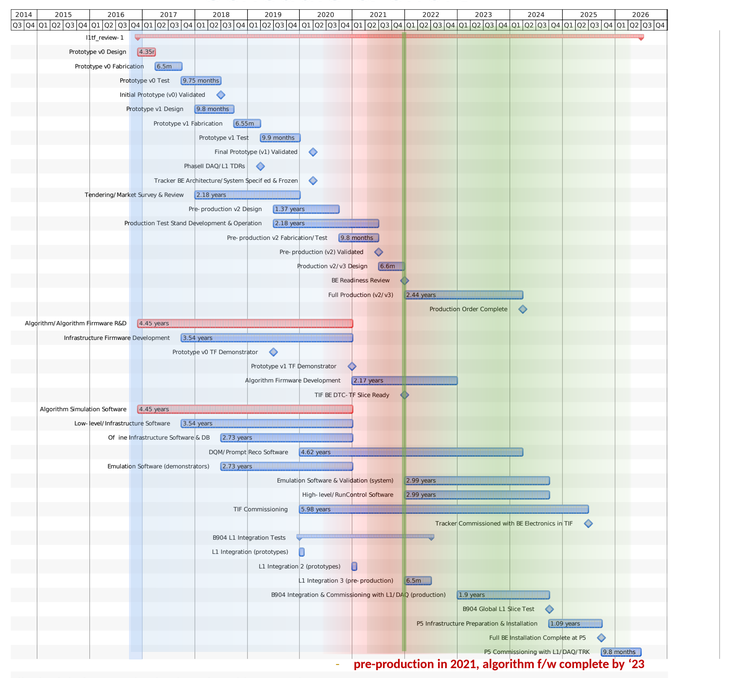
Full system - Summary

	Conservative	Baseline	Advanced
Link Speed	10 Gbps	16 Gbps	25 Gbps
T	36 BX	18 BX	12 BX
$N_{DTC\_links\_out}$	72	36	24
$N_{TFP\_links\_in}$	64	64	64
$N_{TFP\_links\_out}$	8	4	2-3
N_{TFPs}	288 (8x36)	144 (8x18)	96 (8x12)
Poduction Cost	6.9 MCHF	4.3 MCHF	?

conclusion - use 'baseline' as model for schedule

- difficult to give better appraisal of 25G feasibility/cost option today
- it is expected cost will be similar/greater than 'baseline scenario' just described, but with little benefit at this stage
- but 25G should be kept in mind as R&D enters next phase, and prototyping planned accordingly

Schedule & effort



- pre-production in 2021, algorithm f/w complete by '23

overall schedule for a production based on this system is compatible with both Tracker and L1 schedules

expect parallel algorithm development with prototype hardware development till 2020/21

- two cycles of prototyping (e.g. links, FPGAs..)
- 3-4y to define, refine, improve algorithms & f/w
- focus on infrastructure firmware/software effort
- proto-demonstrator milestones ('19, '21)
- adaptation to system constraints (e.g. L1 requirements/interfaces, DTC/cabling spec, DAQ and common electronics/infrastructure etc.)
  - Hardware 30 SY
  - Firmware 38 SY
  - Software 40 SY
- production test-stand definition/development
- preparation for integration & commissioning; slices in operation at TIF (TK & DTC) and B904 (L1/cDAQ)
- peak of software effort

- **BE readiness review early 2022 to green light production**
  - production schedule has margin if needed (2.5y)
  - have learnt to expect the unexpected...
- **integration & commissioning with L1 systems, DAQ, DTCs and Tracker begins early**
  - required to guarantee tight P5 installation schedule
  - expect regular month long L1 integration tests starting ~2020/21

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n° [317446] INFIERI "INtelligent Fast Interconnected and Efficient Devices for Frontier Exploitation in Research and Industry"