Imperial College London

The LHC and the really big data challenge

Instrumentation course guest lecture

Andrew W. Rose, Imperial College London awr01@imperial.ac.uk

I'll upload a copy of these slides to www.hep.ph.ic.ac.uk/~awr01



0

Introduction

 I hope you will all be familiar to some extent with the

Large Hadron Collider

at CERN, Switzerland but will assume minimal background knowledge

- The largest and most complex scientific endeavour in history
- Most famous for its observation of the Higgs boson

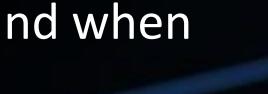


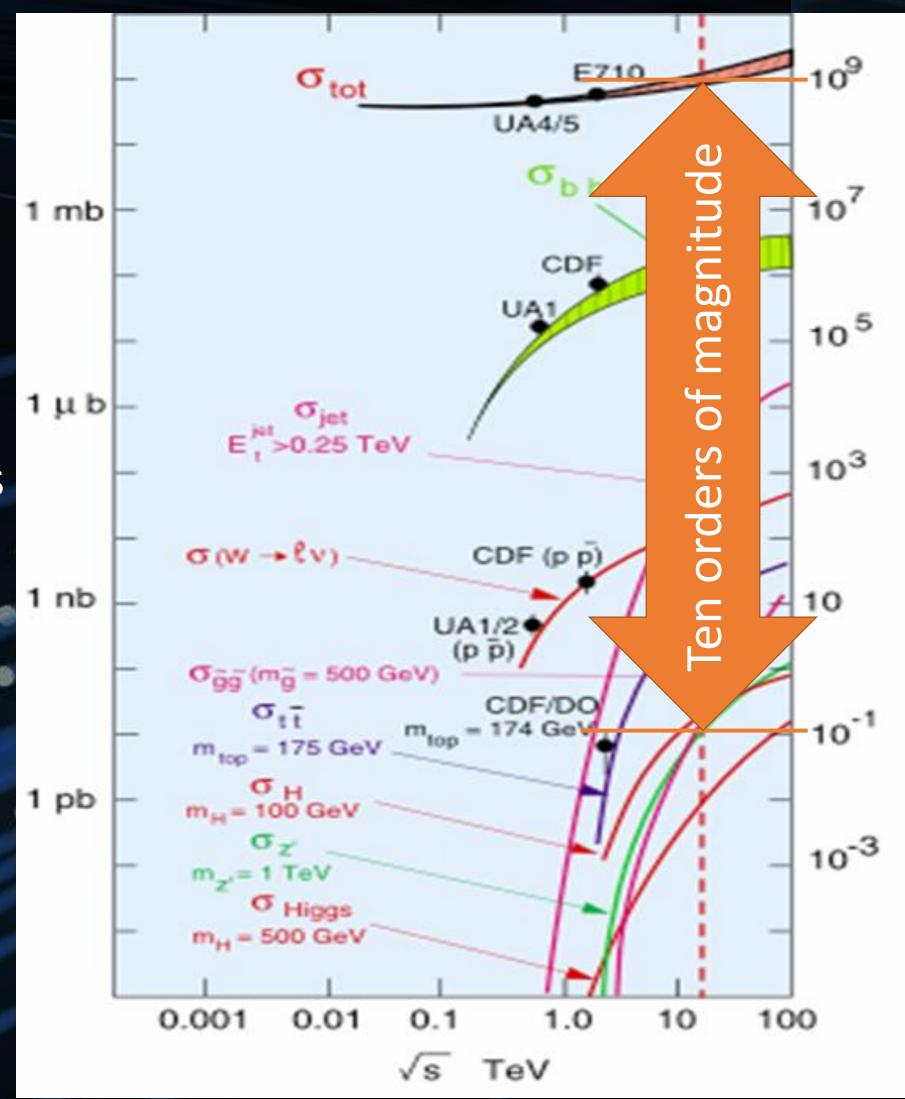
• A bold statement:

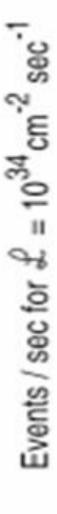
Science is the art of knowing what to record, and when



- Science is the art of knowing what to record, and when
- With CMS & ATLAS in "discovery mode", we care about the Higgs ightarrowBoson or rarer
 - Higgs Boson production is ten orders of magnitude below the total ightarrowinteraction rate







- Science is the art of knowing what to record, and when
- With CMS & ATLAS in "discovery mode", we care about the Higgs Boson or rarer
 - Higgs Boson production is ten orders of magnitude below the total ightarrowinteraction rate
 - That is a needle in a haystack the same mass as the **Empire State Building**





- Science is the art of knowing what to record, and when
- With CMS & ATLAS in "discovery mode", we care about the Higgs ightarrowBoson or rarer
 - Higgs Boson production is ten orders of magnitude below the total \bullet interaction rate
 - That is a needle in a haystack the same mass as the **Empire State Building**
 - Or if you collided protons once per second, that is one event every







- Science is the art of knowing what to record, and when
- Boson or rarer
 - Hig s We need loads of them to be able to study them... ulletinteraction rate
 - That is a needle in a haystack the same mass as the **Empire State Building**
 - Or if you collided protons once per second, that is one event every





And one Higgs boson is pretty much With CIAS & ATLAS in "discovery mode", we care about the Higgs USELESS





~50 protons at a time X 40 million times a second



1 Higgs boson every ~5 seconds

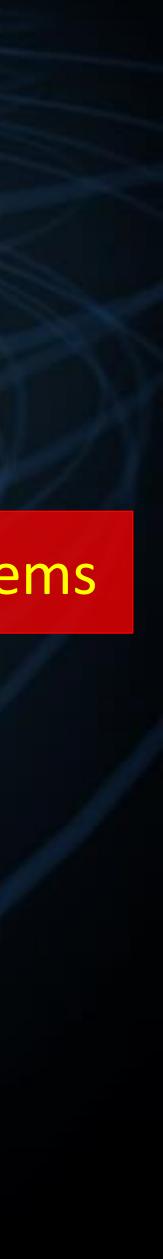
~50 protons at a time X 40 million times a second



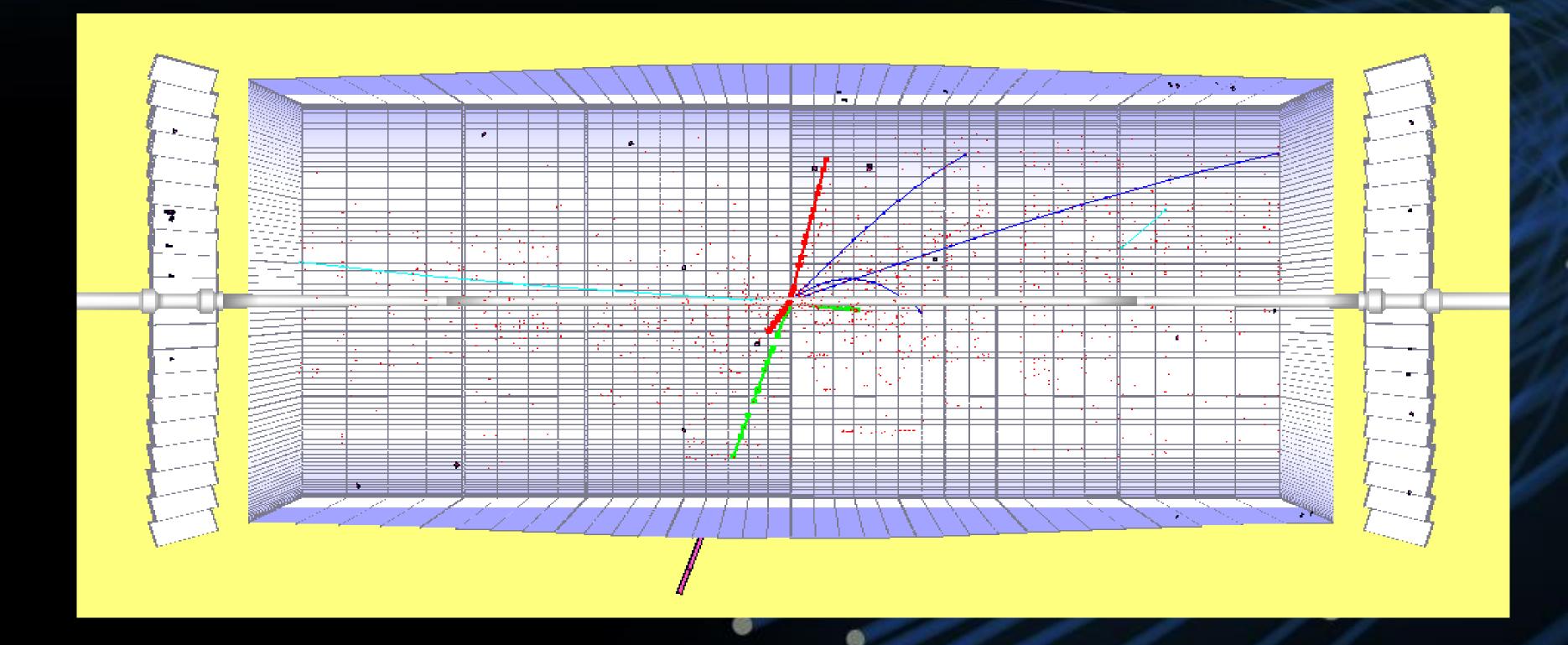
~50 protons at a time X 40 million times a second

1 Higgs boson every ~5 seconds

But this causes problems

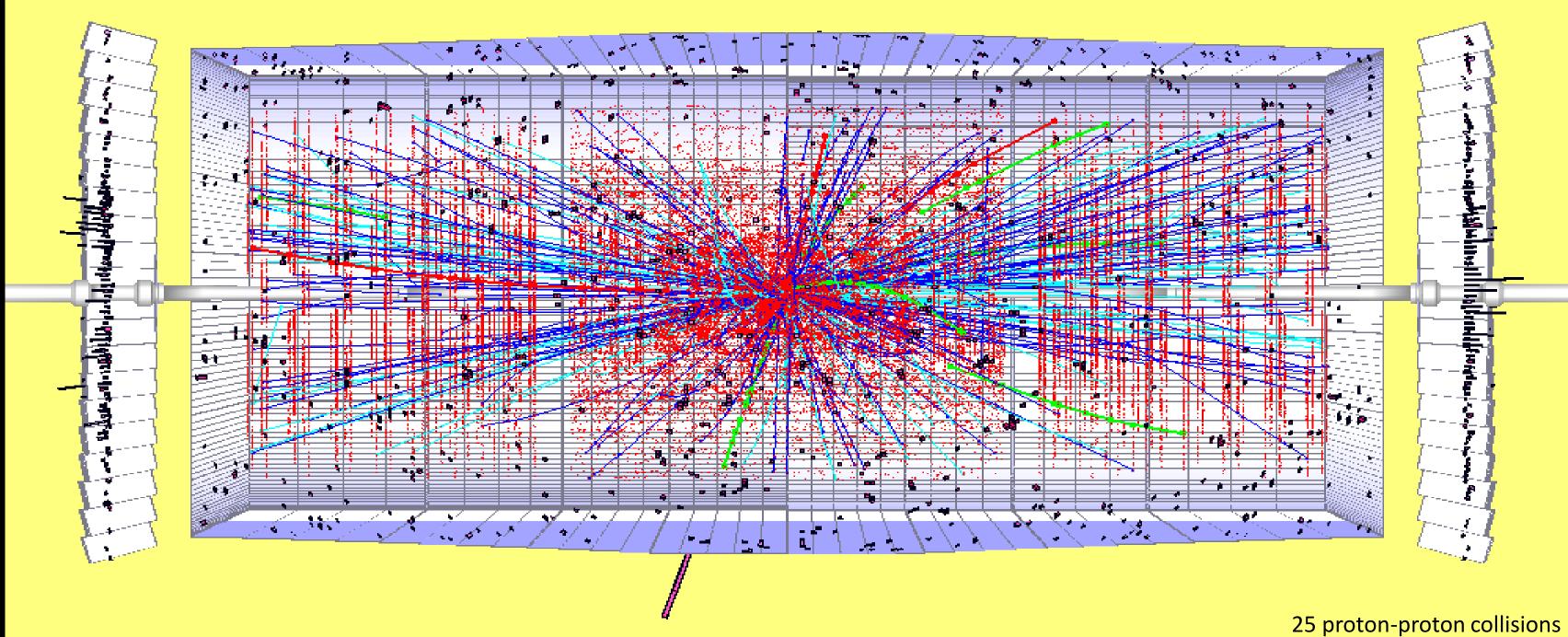


One proton-proton collision



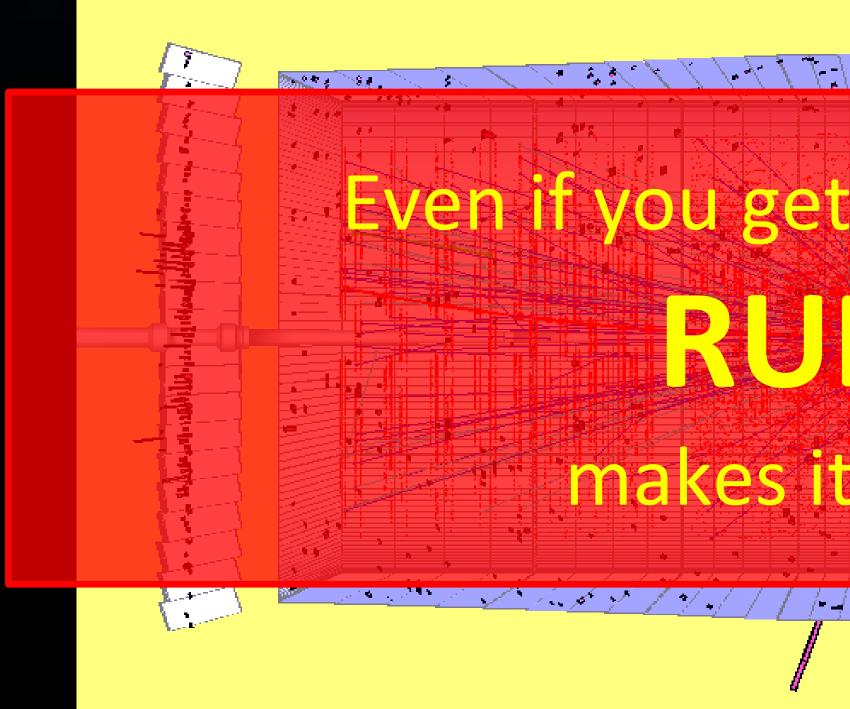


Many proton-proton collisions





Many proton-proton collisions



Even if you get a Higgs boson, the **RUBBISH** makes it hard to find

25 proton-proton collisions



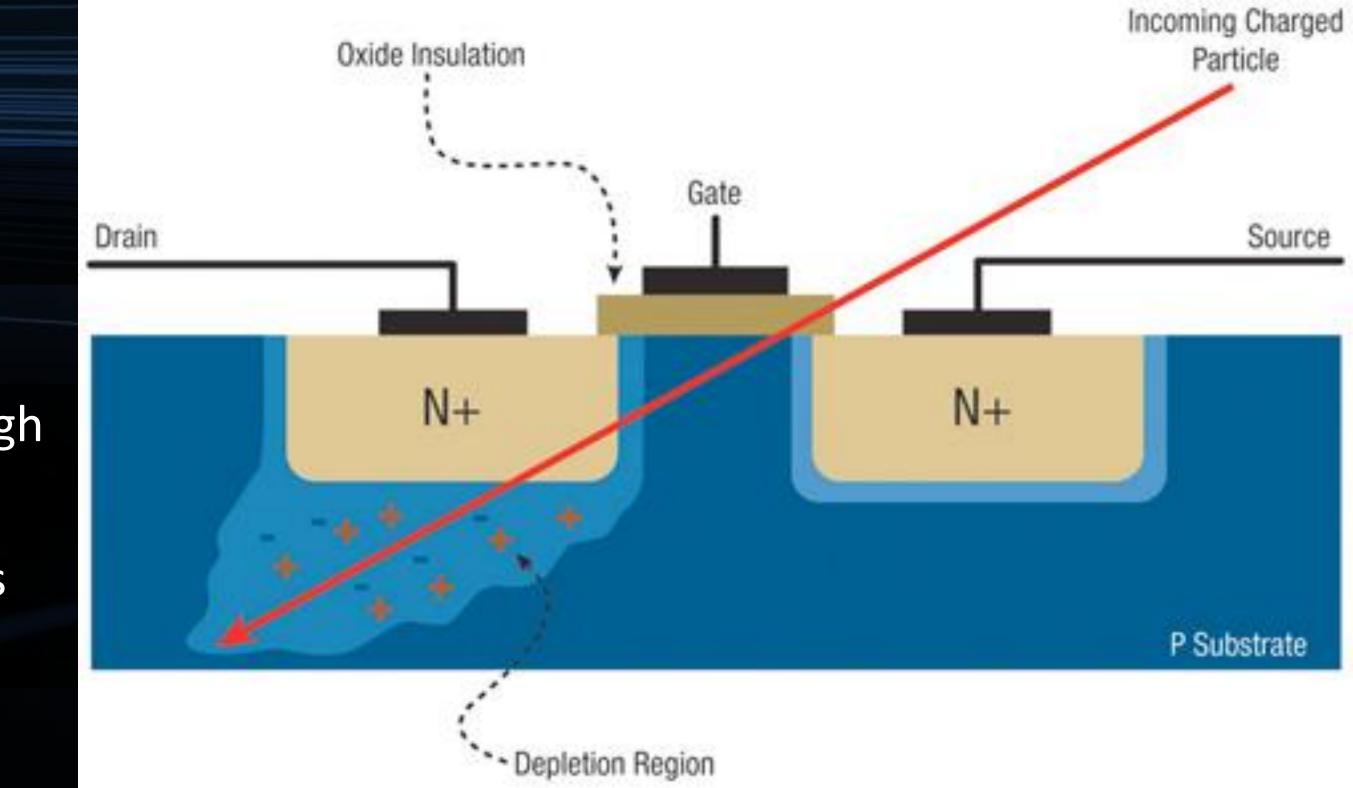
1 Higgs boson every ~5 seconds

~50 protons at a time X 40 million times a second

This causes other problems too

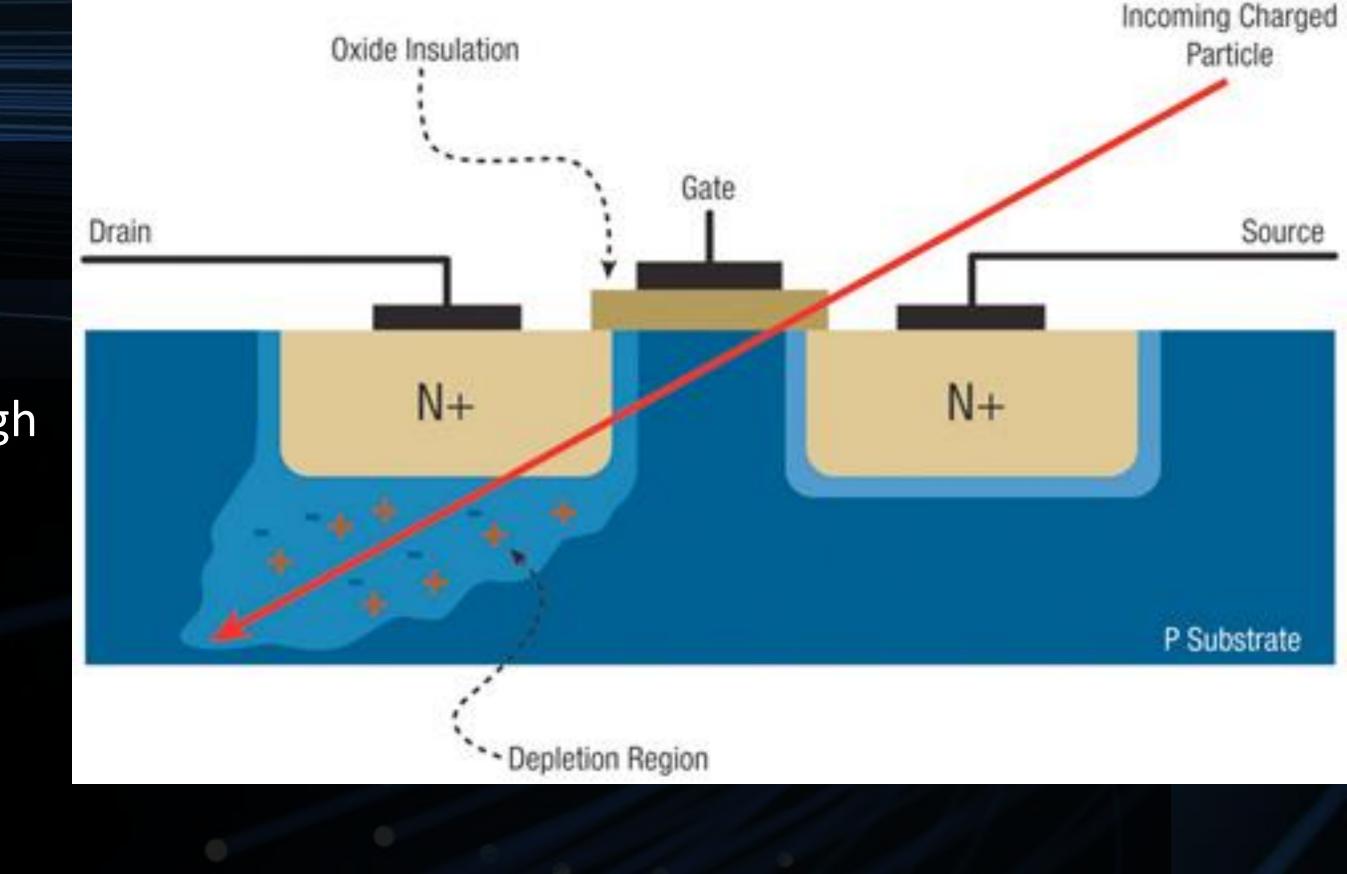


- When energetic charged particles pass through a transistor, they can change the state
 - And there are plenty of charged particles in the LHC experiments



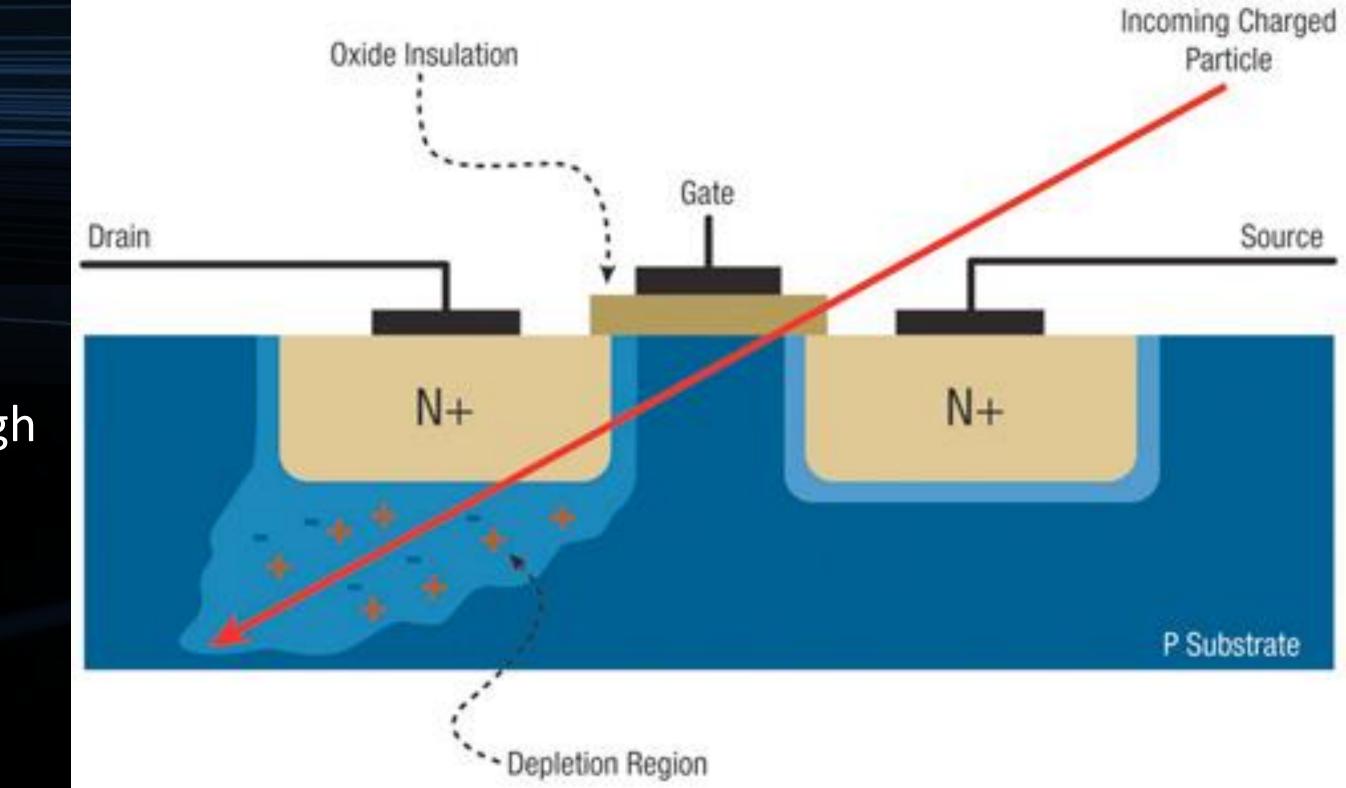


- When energetic charged particles pass through a transistor, they can change the state
 - And there are plenty of charged particles in the LHC experiments
- In a microprocessor
 - At best this corrupts the data
 - At worst it changes the program flow





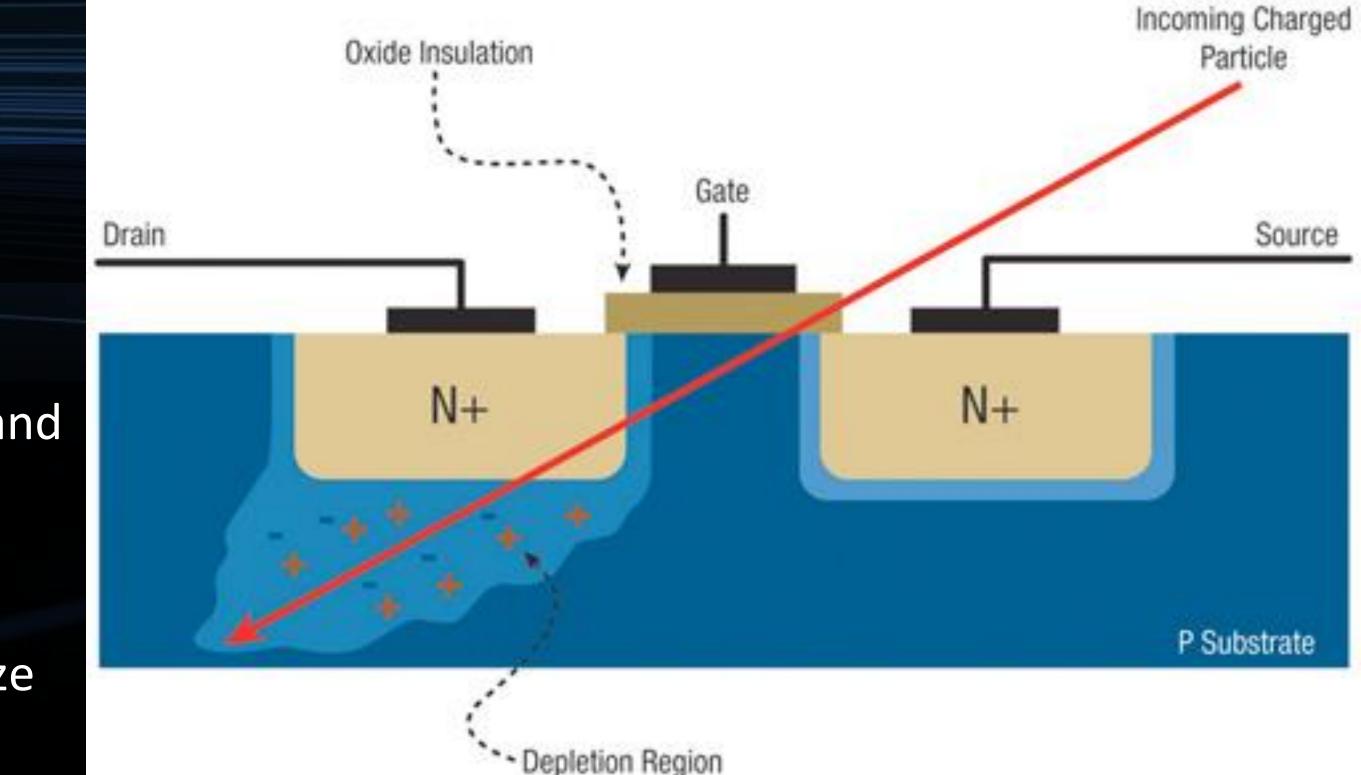
- When energetic charged particles pass through a transistor, they can change the state
 - And there are plenty of charged particles in the LHC experiments
- In a microprocessor
 - At best this corrupts the data
 - At worst it changes the program flow
- There are much, much worse effects
 - Entire circuits reconfigured
 - Transistor shorts power to ground burn out the chip
 - Block the ability to reset or reconfigure the chip



n out the chip he chip



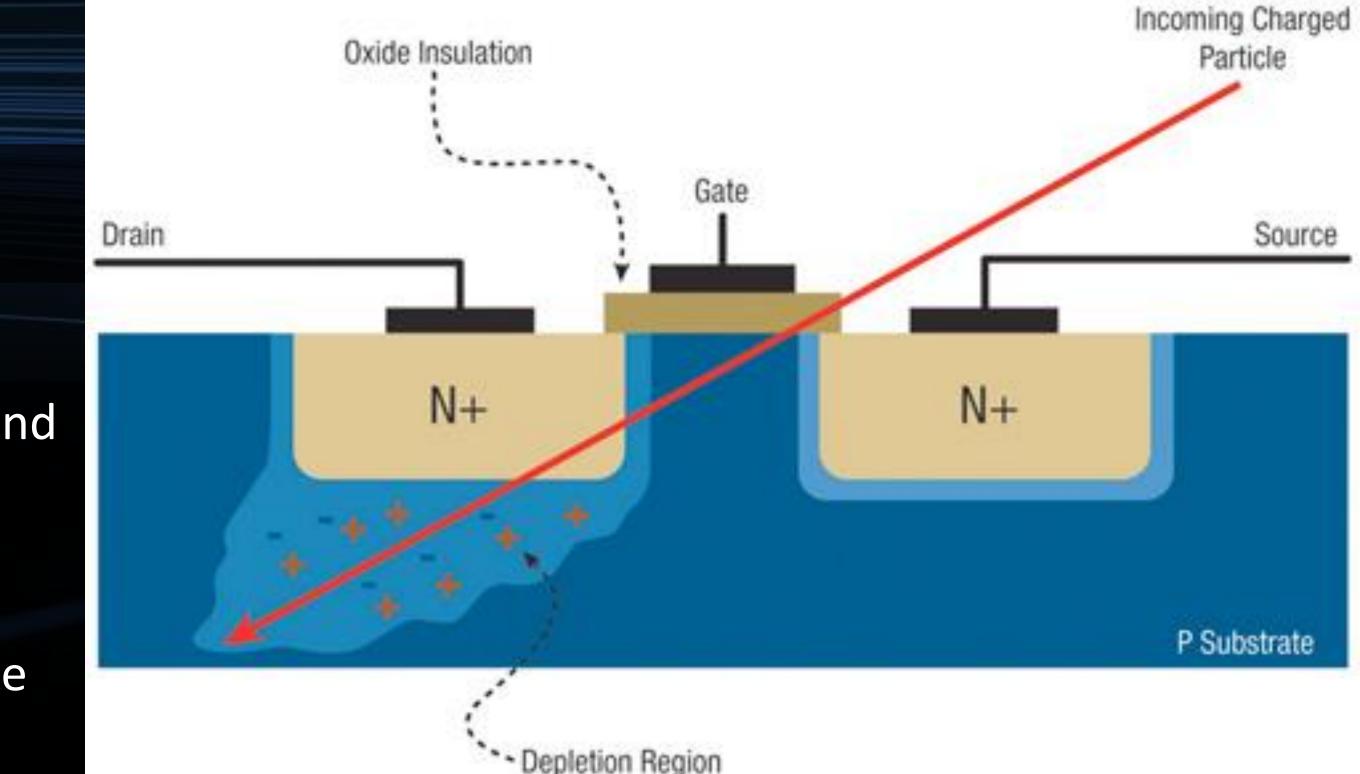
- Mitigate by having three copies of the logic and ightarrowarbitrate by majority logic
 - Triple Redundant Logic
- Deep well transistor architectures to minimize \bullet charge collection
- Insulating substrates rather than semiconductors
 - Diamond, Sapphire, ... ullet
- Or wide band gap substrates ightarrow
 - Silicon Carbide, Gallium Nitride
- **Continual reconfiguration**
 - Called "Scrubbing"



Designing radiation-hard electronics is a specialist skill



- Mitigate by having three copies of the logic and ightarrowarbitrate by majority logic
 - Triple Redundant Logic ightarrow
- Deep well transistor architectures to minimize \bullet charge collection
- Insulating substrates rather than semiconductors 0
 - Diamond, Sapphire, ... ullet
- Or wide band gap substrates ightarrow
 - Silicon Carbide, Gallium Nitride
- **Continual reconfiguration**
 - Called "Scrubbing"

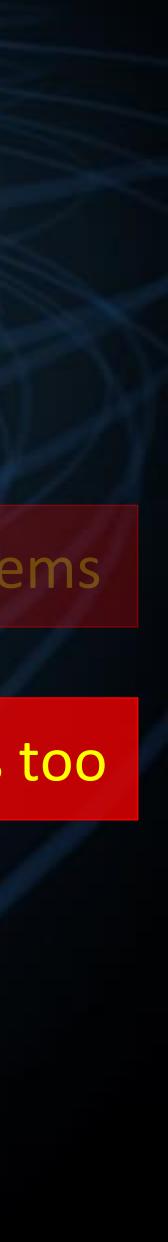


The best solution is to keep your most sensitive electronics as far away from radiation as possible



1 Higgs boson every ~5 seconds

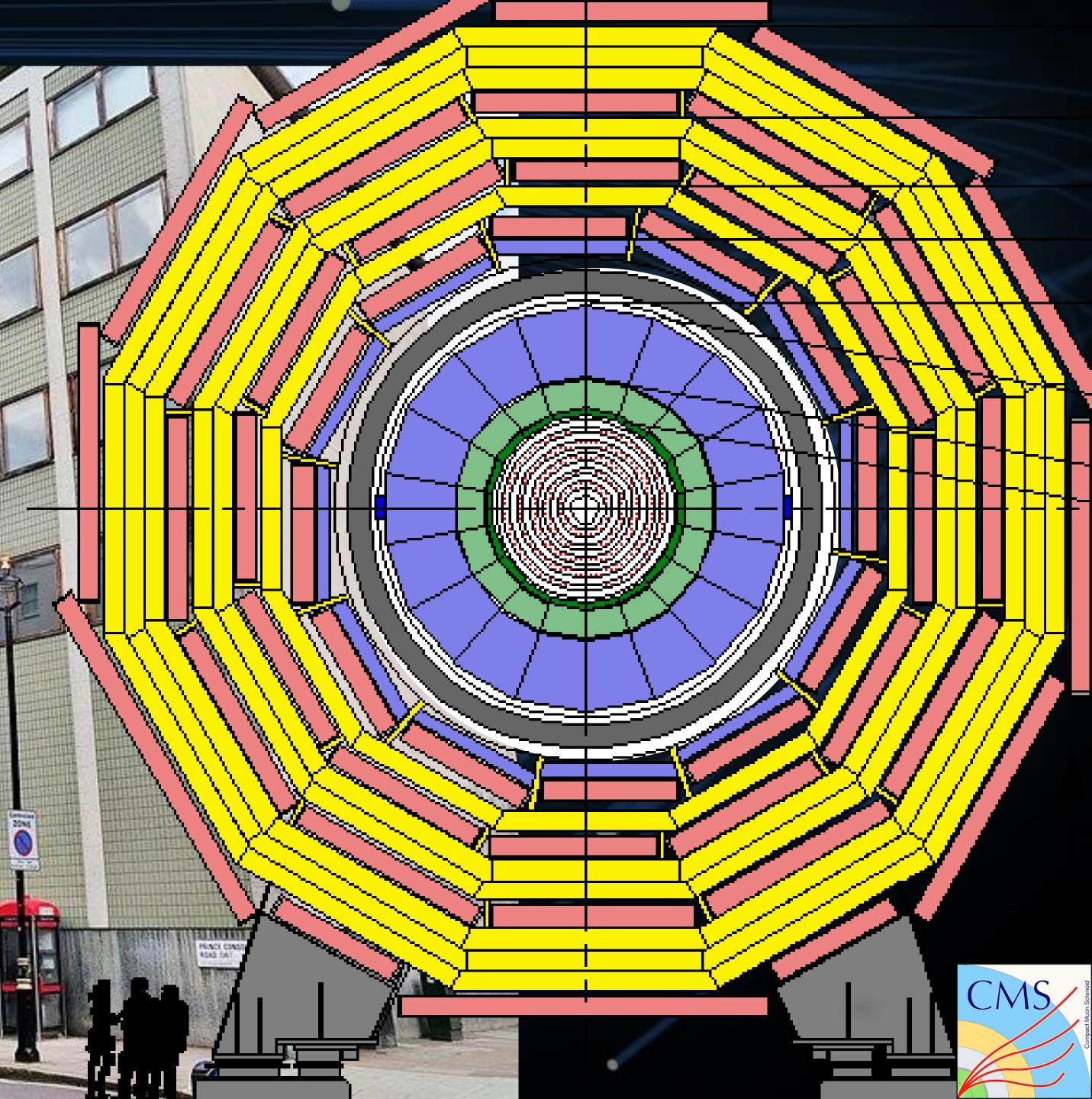
~50 protons at a time But this causes problems X 40 million times a second This causes problems too



Big detectors for small particles

11

係



Big detectors...

~65M Silicon Pixels

~10M Silicon Strips

~80k PbWO₄ Ecal Crystals

~15k channel Brass/Plastic sampling HCAL

~568k RPC/DT/CSC Muon channels



Big detectors...

× 40 million measurements per second

~65M Silicon Pixels

~10M Silicon Strips

~80k PbWO₄ Ecal Crystals

~15k channel Brass/Plastic sampling HCAL

~568k RPC/DT/CSC Muon channels



... Bigger data

× 40 million measurements per second

~65M Silicon Pixels ≡ 21 PBit per second

~10M Silicon Strips
≡ 4 PBit per second

~80k PbWO₄ Ecal Crystals \equiv 40 TBit per second

~15k channel Brass/Plastic sampling HCAL≡ 10 TBit per second

~568k RPC/DT/CSC Muon channels \equiv 23 TBit per second



... Bigger data

What does that even mean?

~65M Silicon Pixels ≡ 21 PBit per second

~10M Silicon Strips≡ 4 PBit per second

~80k PbWO₄ Ecal Crystals \equiv 40 TBit per second

~15k channel Brass/Plastic sampling HCAL≡ 10 TBit per second

~568k RPC/DT/CSC Muon channels ≡ 23 TBit per second



What does that even mean?

1 Exabyte

1,000 Petabytes or 250 Million DVDs



400 Exabytes

The amount of data that crossed the Internet in 2012 alone

100 Exabytes

A video recording of all the meetings that took place last year across the world

5 Exabytes

A text transcript of all words ever spoken[†]

100 Petabytes

The amount of data produced in a single minute by the particle collider at CERN

480 Terabytes

×200

A digital library of all the world's catalogued books in all languages

https://www.cisco.com/

1 Yottabyte

1,000 Zettabytes or 250 Trillion DVDs

1 Zettabyte 1,000 Exabytes or 250 Billion DVDs

20 Yottabytes

A holographic snapshot of the earth's surface

300 Zettabytes

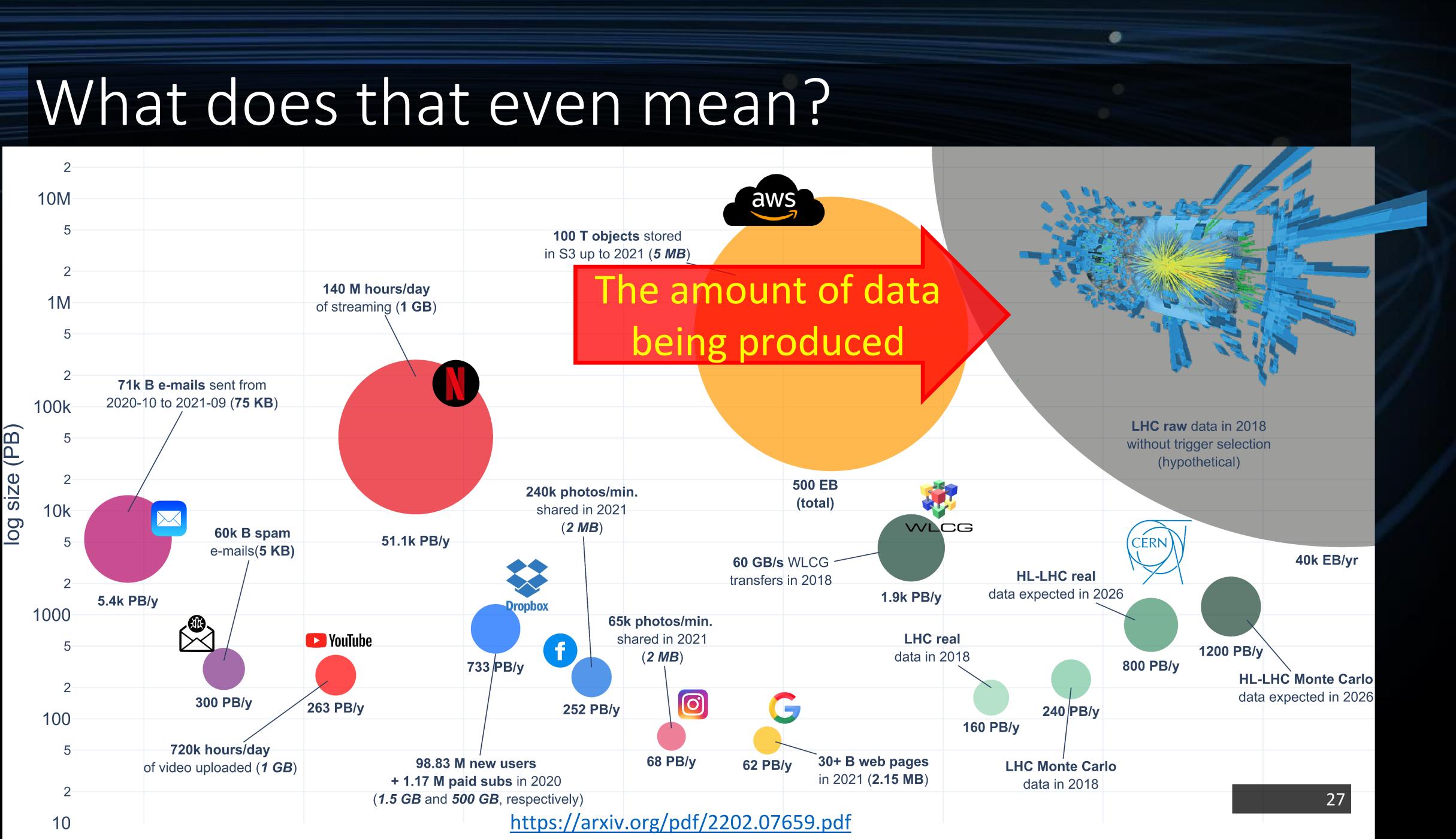
The amount of visual information conveyed from the eyes to the brain of the entire human race in a single year

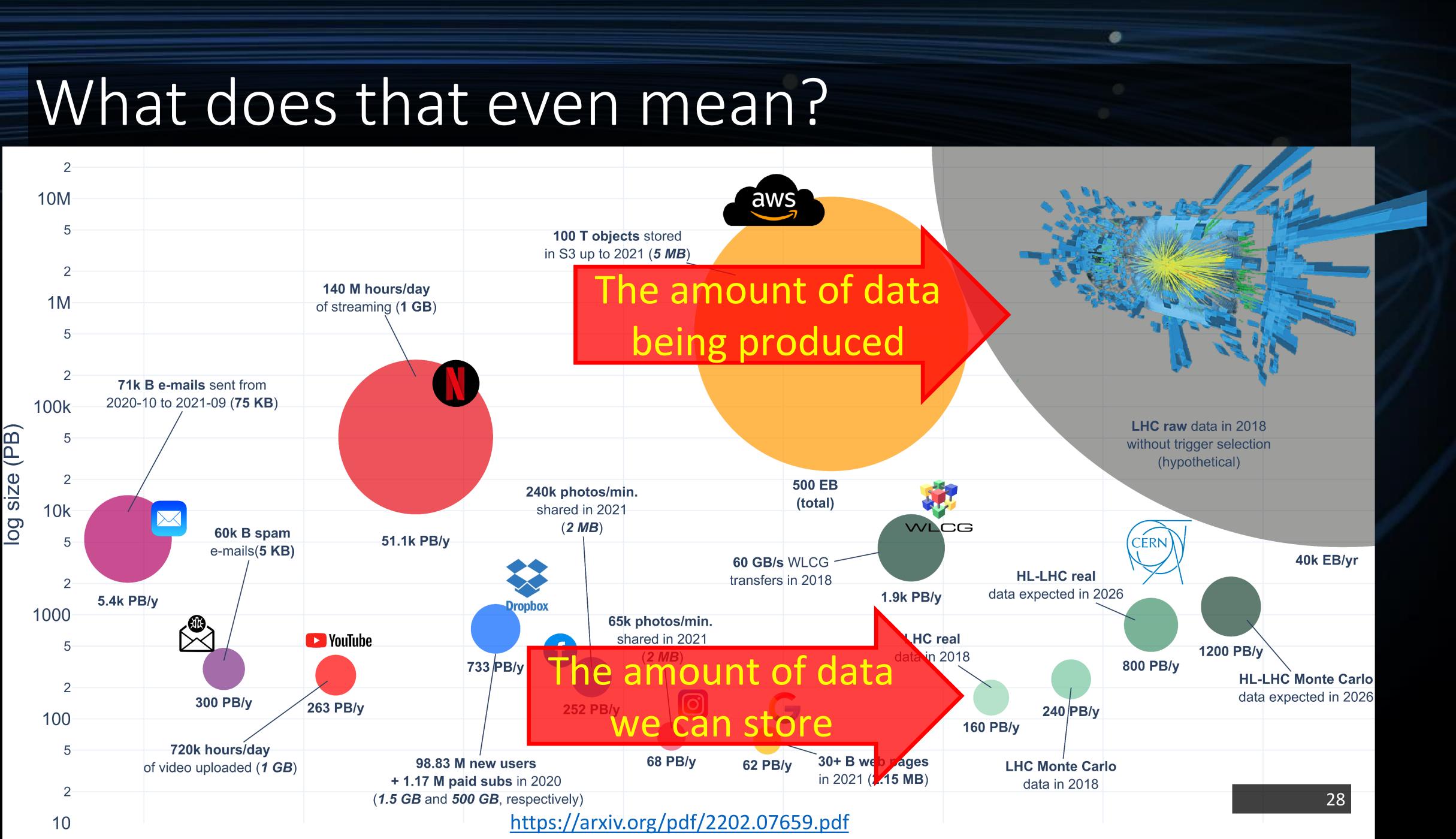
1 Zettabyte

The amount of data that has traversed the Internet since its creation





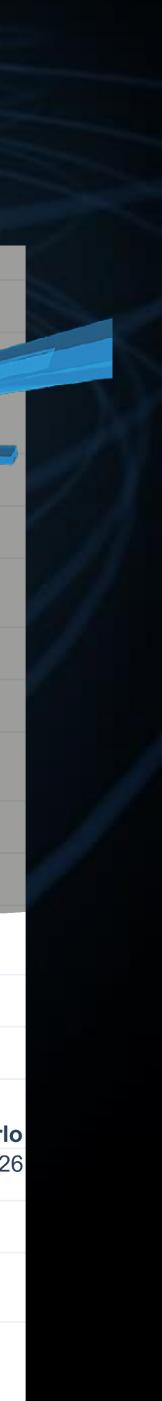




What does that even mean?



40k EB/yr 1200 PB/y **HL-LHC Monte Carlo** data expected in 2026

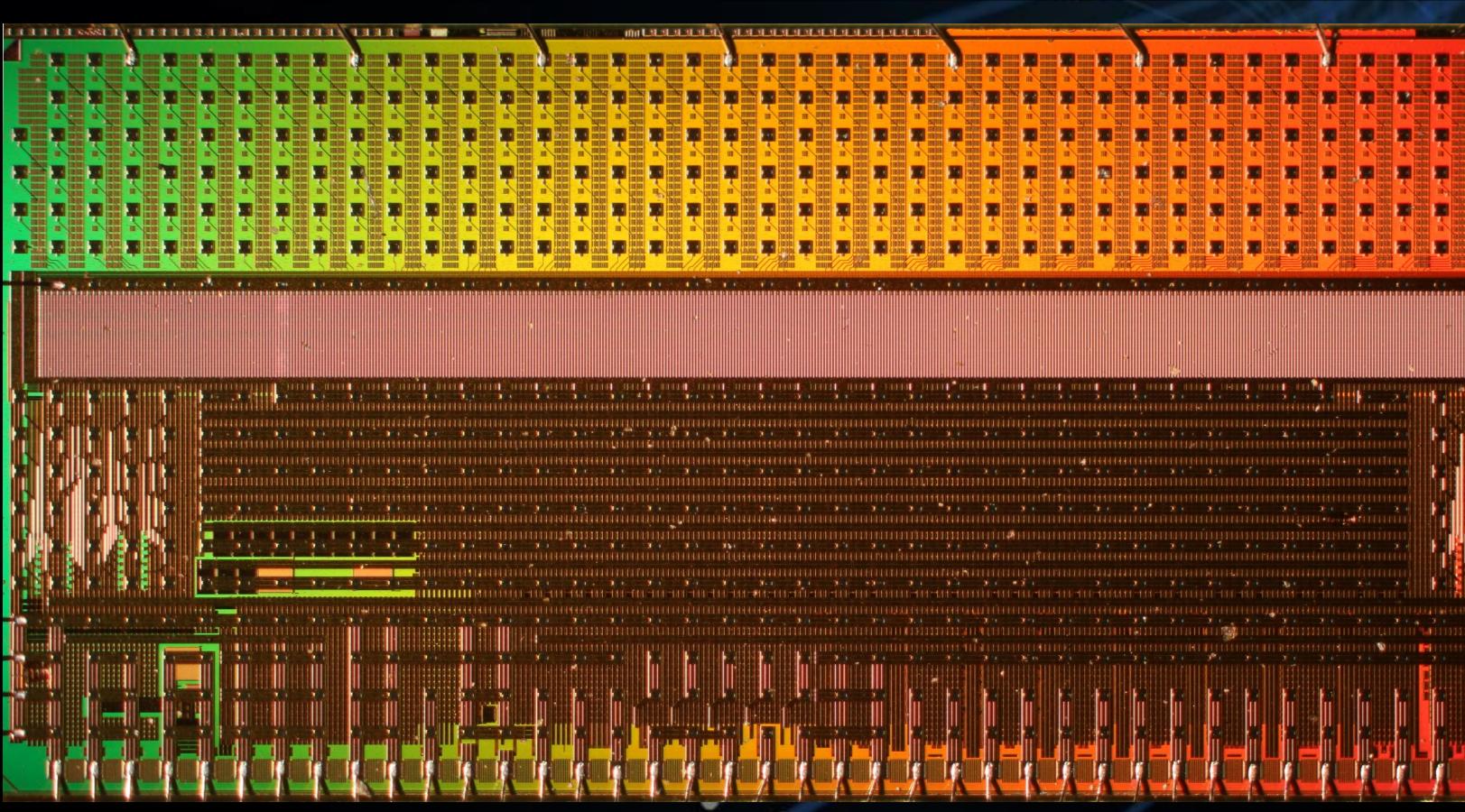


We keep the data safe on the detector



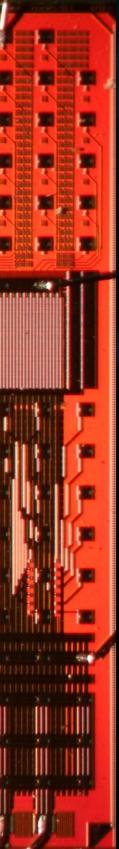
• We keep the data safe on the detector

CMS CBC3 ASIC Designed at Imperial College London 🐰 Layout at Rutherford Appleton Lab



Designing radiation-hard electronics is a specialist skill

		a statement of the stat			THE R. L.																-					
NUMBER OF STREET	11 C	N HHILS STREET	in the second	ष प प प	1111	1111	11-11-11	H. H. H.	N. N. N.	11 11 11	- Instrum				yna			-	Provide and	and the second	and	10, 1000				Alterna 1
					the second s							A REAL PROPERTY AND A REAL														
-				-					-		-	A summing in		-		-			-	-			-	-	and second as	
a terrare	AND ADDRESS AND ADDRESS	The second second	TRADES TO STRATES	L	COLUMN STREET	and an and a state of the local division of	AL DESIGNATION OF	a minima)	and a second second	A DESCRIPTION	THE PARTY OF	TITUER .		Contraction of the local division of the loc	TRADUCTOR OF		TRAFILE A	COLUMN STREET,	TELEVILLE	A DESCRIPTION OF	TRANSPORT	ALC: NO.			and the second	
I IIIIII	A BINADA A BURN	I T INNI T		A MILLING	C INTINA	C IIIIIII	IN COLUMN 2	E HIMM	C milita	T IIIIII	A DIMME		-	A HINNES	A INNIA	THEFT	TININ .	1		· INISIA	4 HIME	A SIMIL	A BILLING	A DISCO	A. STREET	A BREAK T
E IN PRIMITS	III (IIIIIII III (IIIIII	I III VIIIIII SII V	INNER IN VIENE	III FIRINI	III (IIIIIIII II	a finne a	I VIINING II	a fainning i	IN THE OWNER	III (IIIIIII	III FIININ	III FIIIIII	III (IIIIII	HR FHIMME		III FILMIN I	III FILING		III PRIMI	HA PHINHA	HI STINGT	III ISABIN		III MININ	III JUININ	11 (11111) II
I INCOM	ARTIN ARTI	I INCIDE I	Rinn Winn	ENGINE .	(MALINE	CHALLER	ANALISE	ANUIR	BRAIN	Contraction of the second seco	Contraction of the second seco	SHIII S	CHARLEN IN CONTRACT OF CONTRACT.	CHARLEN .	CALCULAR OF A	CHINE .	antin	CHANNEL .	COMPLETE IN COMPLETE INCOMPLETE IN COMPLETE IN COMPLETE IN COMPLETE IN COMPLETE IN COMPLETE INCOMPLETE INCOMPLIANTE INCOMPLICATI INCOMPLICATI INCOMPLIANTE INCOMPLIANTE INO	SMILLIN .	CHARGE STREET	Saliting	Suppose	Statute of	Station of the	AND IN CO.
COLUMN COLUMN				Contraction of the local division of the loc	THEFT.	TITLE A	THE REAL PROPERTY AND INCOME.	COLUMN T	THE PARTY OF		and another	THE PARTY OF	STATES	Title	TITUTE I	And strend in		TITLE OF	TRADUCT	TITLE I		TRANSPORT	ALC: NO.			
-	A DESCRIPTION A DESCRIPTION	a straffe a		A DISTRICT	4 1912010	INTERE	INCOME.	INCOME.	CITERIN .	In the second	- IIIIII	4 mette	-	THE REAL	-	4 111111	a minana	A SIMILA	-	A INIBIA	4 1612512	4 1318518	A BELLEVIL	4 - 1568.8	A DISTRICT	A BIRD A
IN INCOME	III IIIIIII III IIIIII	III LINNIN III]	IIIII III IIIIII	III IIIIII	III	I IIIIII II	anna i	B BIBBIB ;	IN INTERIOR	III IIIIII	HI HING	ITS ITHEFTE	SH HINNI	HE DING	III IIIII	ILS INCOME I		III IIIIII	HR SHARE			HI STING	III IIIII	III IIIIII	IN ILLING	
No. of Lot of Lo		NUM	State Name	SUPPLY A	SUPID	CALCULATION OF THE OWNER OWNER OWNER OF THE OWNER OWNE	Number of Street	NUMBER OF	NUT III	NUT III	SUPIE I	CLUTTE .	Supple	SUPPLE -	COLUMN TO A	Surfie .	The state of the s	Contract of the local division of the local	No. of Concession, Name	No. 11	Number of	Care a	Caperie .	City in a	TOTAL COLUMN	Sugar .
10000			122312	TITLET A	1111111	annorth - bea	1111111	-	and and a second	TRAILS.	and attant	110000			10000	Carl attacts	1 111111		100000		and an other the	110010	TAXABLE INC.		STREET, STREET, ST	
110 110	100 010 1 100 01	110 110	12 010 100	110.010	1 112 112	III III	100 010	120 012	III IIII	III III	118.118	110 110	III III	-	100000	110010	HEAL	1 田田	1 11 11		1 110010	a anglan	a min	Bullet a	THE PARTY	Tables 1
HAR BEE ME	III IIIIIII III IIIIIII			III IIIIII	III IIII III II	I III III II		HE HE HE	III III	HE HERE	III IIIII	III IIBIIIB	III IIIIII	IIB IIBBIE	118 ING 118	110 1100110 1	18 18 19 19	III IIIIII	HE HE HE	HI HANNEL	III IIIII			HI HANN	III IIIII	
100.000				1 10 200	110 010	110 010	110 510	110,010	110 010	120 018	118418	1 IN STR	TRACE	110010	TTO STR.	TRACE	O TIMOTA	10010	110 DIST.	10000	1 III STR.	100 BOR	TIR STO	1 California	Tin Tree	Trains -
1000	\$			1000	111 III III				11000		1 11 11 11 11 11 11 11 11 11 11 11 11 1	TRACES.		TRACE	11000	TENTE N	TRAIL I		THE REAL PROPERTY.	TRACE IN CASE OF			TRADE	TRACE I	TRANSPORT OF	1011
	1000			110 010	1 11 111	110 110	110 110	110 010	III III	111 111	110 318	110 110	110010	10010	118.00	III III	HEAL	118118	in its	11811	110111	THE HEAD	田田田	IN THE OWNER	THE STREET	100100
								I III III		118 118 118	110 110 110	ISB IN THE	HR HIRES	III IIIII	110 11011	III IIIII		III IIIII			HI INCOME	III IIIII		HI HINK	HALL HALL	THE PORTON IN
The star		THE REAL PROPERTY AND INCOMENTS		TINGER .		1 CON STR.	1100000	LTB BEB		10000	110 110	120 212	1 COLORED	C IDEN	TRATE.	T Da Sta	TORNES .	Canal State	A DECEMBER OF	Tan Ita	The second	1.4	Lange and	No.	a state	Taking .
			main ann an an an				an an an an	In Inter 1	11 III III I		Tent Inches	1200		THE REAL			1 124	18 HE HE	Hall Hall	THE R. LANSING MICH.	11111		18 H H	The sector 1	Inter 1	and inside the
	11111 11111			and in the second					田田		and its its	112 HE		11000	1999		田田田	日田田	田田田	THE PARTY IN COLUMN	and this	1898	田油	日日日	田柏	In the second second
111 111 111	III TANAN - III TANAN		IN SAN DEN SYN AND	III IIIIII	ITH FYRIAM I	10 1 10 10 10 10 10		110 010 1	THE REAL	III IIIIII	III IIIIII	tan ten ate	TER PERMAN	THE TRACE	THE PERSON	ILE ITERSTE	THE STREET	178 118 118	ILE ITERIT	ILS LINKS	1130 1120 1120	Hand Hand	HI LOUDA	THE STREET	THE DESIGNATION OF	THE FRENCH HA
				THE REAL	Die ste	1 III III	THE REAL	ALC: NO.	THE REAL	- Imain	1 milde	THE REAL	Take.	THE REAL	Ta ita	Takis .	hand	Denter	No.	Datis	Terre	The	Nash	Taking .	Neth	Destin
							TRANK W	I I I I I I I I I I I I I I I I I I I			THE REAL PROPERTY AND			THE REAL PROPERTY AND INCOME.		THE REAL PROPERTY AND INC.					COL HERE			THE PARTY I	Internal Int	and light and
	AND LEASE AND LEASE	Carling Statist	RAILS - LANSING	AND LEADER	Linite a	A INCHIEF		LIB/SIDE	LINIE		AND A STATEL	AND LAND		A STATE	A RATE	AND LINGTR.		CHARLE I	ANH	AND CRAME	Likitik	CRARK (LINH	thank i	LAAR	AND A DESIGN AND
	AND DESCRIPTION DESCRIPTION		TREESE STREESE	TRADUCT /	ATURTURA	HILIT	A HILL	HIRIDA	Altin	HILLING	- HINGING	IIIIIII		III IIII		H H H		ALL DE LE	HIGHLAN	- ALMIN	HIGHL	Autor	THE REAL PROPERTY OF	11000	and a second	
	/ / minin	(/ upon prop	and a property	entities	Contraction		1	and and	Continue	Superior	S.C. Comment	adaded	(Johnson	dinning (1 - Summer	Contract /	- Junea	(gridings /	- Junita	6000		C.C. Strange		CCCCC		



• We keep the data safe on the detector for several microseconds





- We keep the data safe on the detector for several microseconds
- We quickly analyse some small fraction of the data and select the interesting crossings to keep



- We keep the data safe on the detector for several microseconds
- We quickly analyse some small fraction of the data and select the interesting crossings to keep

Called Triggering

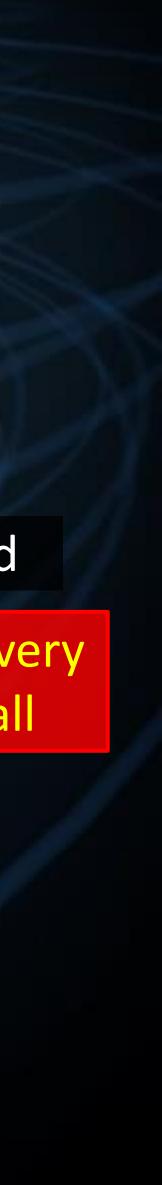


- We keep the data safe on the detector for several microseconds \bullet
- We quickly analyse some small fraction of the data and select the interesting crossings to keep.

uninteresting crossings to discard

The LHC is a discovery machine, after all

Called Triggering



- We keep the data safe on the detector for several microseconds
- We quickly analyse som (small fraction of the data)

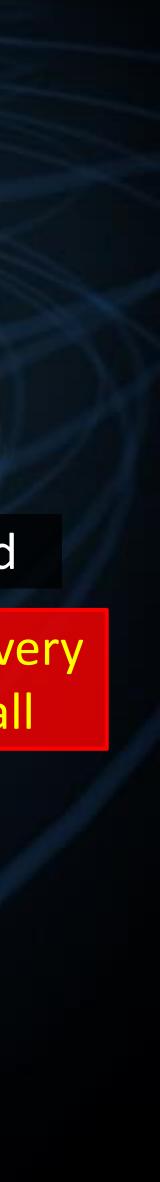
~30Tbps

nd select the interesting crossings to keep

uninteresting crossings to discard

The LHC is a discovery machine, after all

Called Triggering



But!

- If you get it wrong
- If your system fails
- If you take too long

- you throw away your valuable physics



But!

- If you get it wrong
- If your system fails
- If you take too long

-you throw away your valuable physics

And that is a *really, really* expensive mistake



So what do we do?

- We keep the data safe on the detector for several microseconds
- We quickly analyse some small fraction of the data and select the interesting crossings to keep



So what do we do?

- We keep the data safe on the detector for several microseconds
- We quickly analyse some small fraction of the data and select the interesting crossings to keep

But the radiation in the experiment is so intense that the • programmable electronics is kept 100m away The speed of light costs you 1µs to get the data off detector • And 1µs to get the decision back to the detector Leaves 1.2µs for all data processing



So what do we do?

- We keep the data safe on the detector for several microseconds
- We quickly analyse some small fraction of the data and select the interesting crossings to keep

But the radiation in the experiment is so intense that the • programmable electronics is kept 100m away The speed of light costs you 1µs to get the data off detector • And 1µs to get the decision back to the detector Leaves 1.2µs for all data processing

4800 instructions on a 4GHz CPU



But recall...

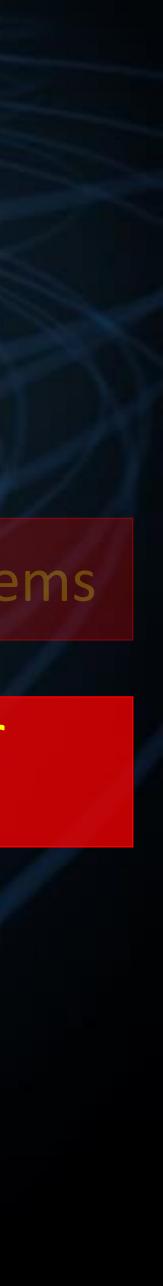
~50 protons at a time X

40 million times a second

But this causes problems

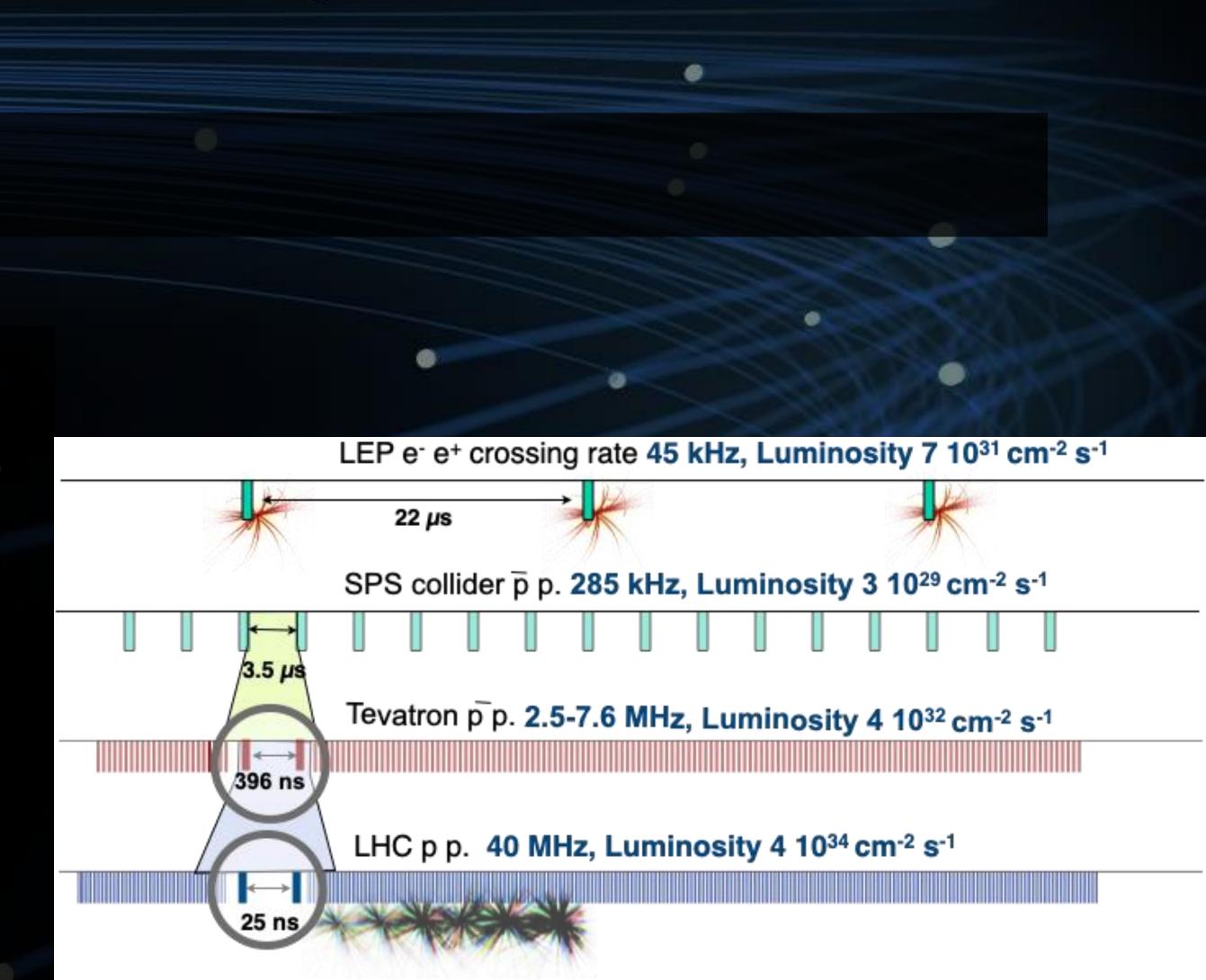
This causes other problems too

1 Higgs boson every ~5 seconds



A note on timescales

- At 40MHz BX rate, a 4GHz CPU could perform 100 CPU operations (not enough to be useful) before the next data arrives
- What technology can we use?



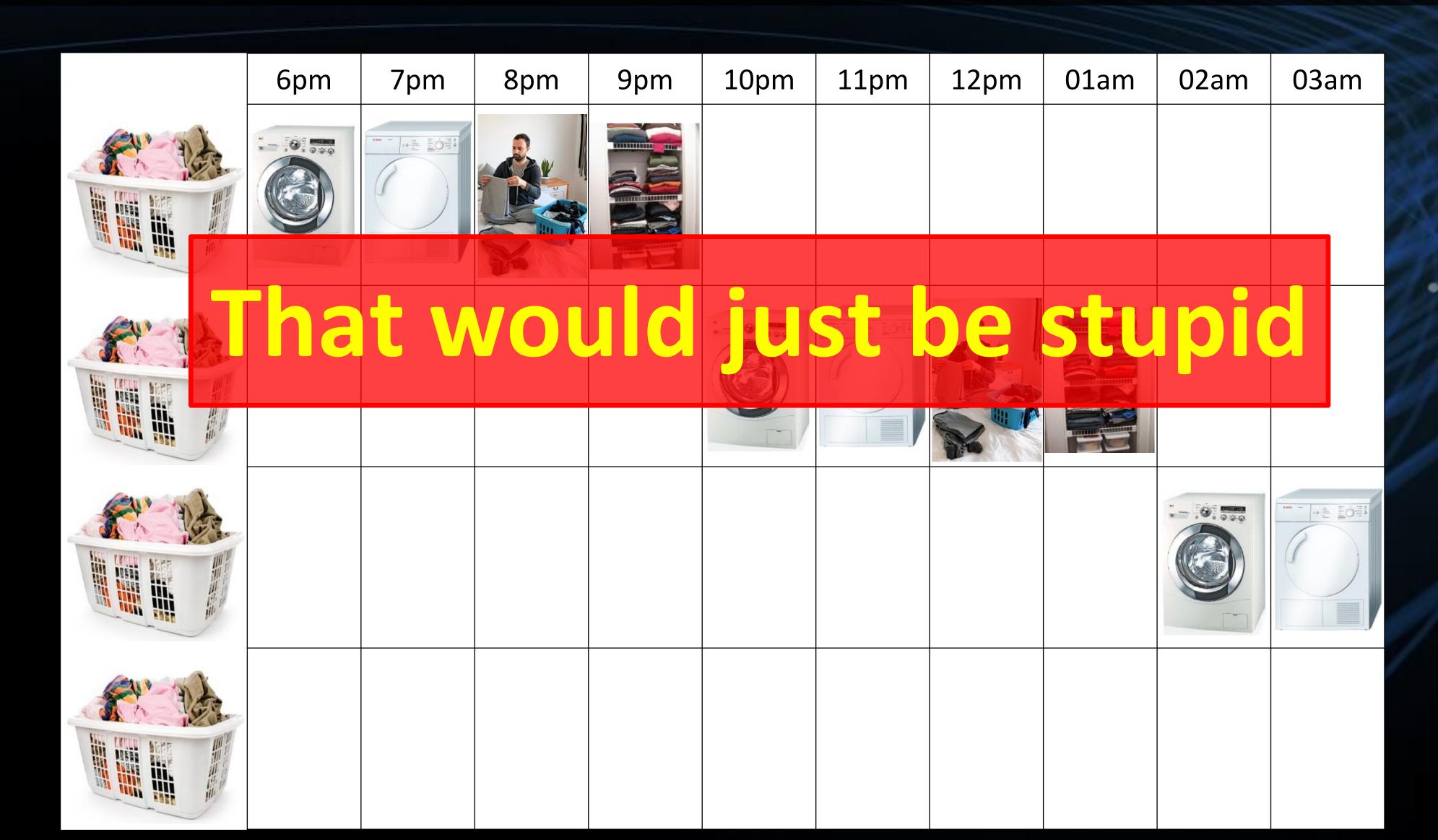


6pm	7pm	8pm	9pm	10pm	11pm	12pm	01am	02am	03am



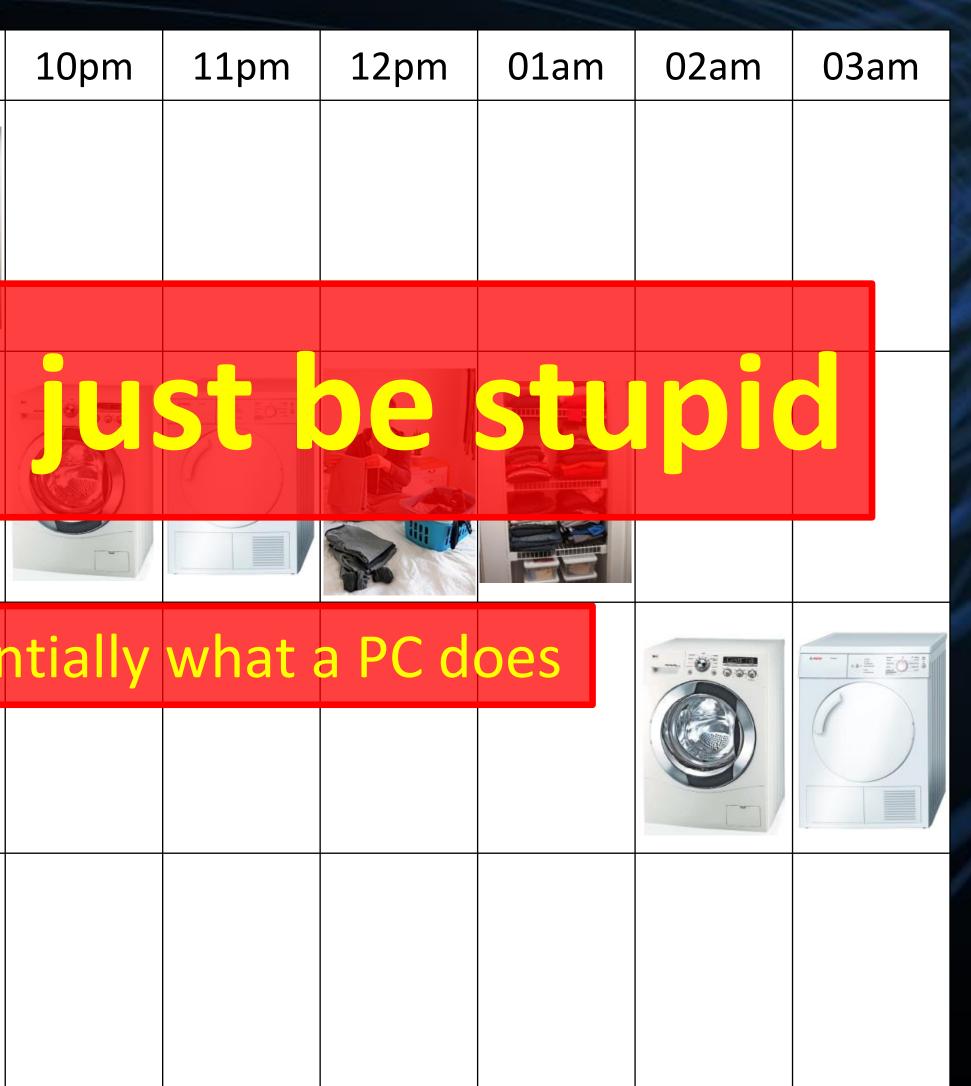
6pm	7pm	8pm	9pm	10pm	11pm	12pm	01am	02am	03am







6pm	7pm	8pm	9pm
h9		101	
	Durf	- thic is	
	DUI		s esser





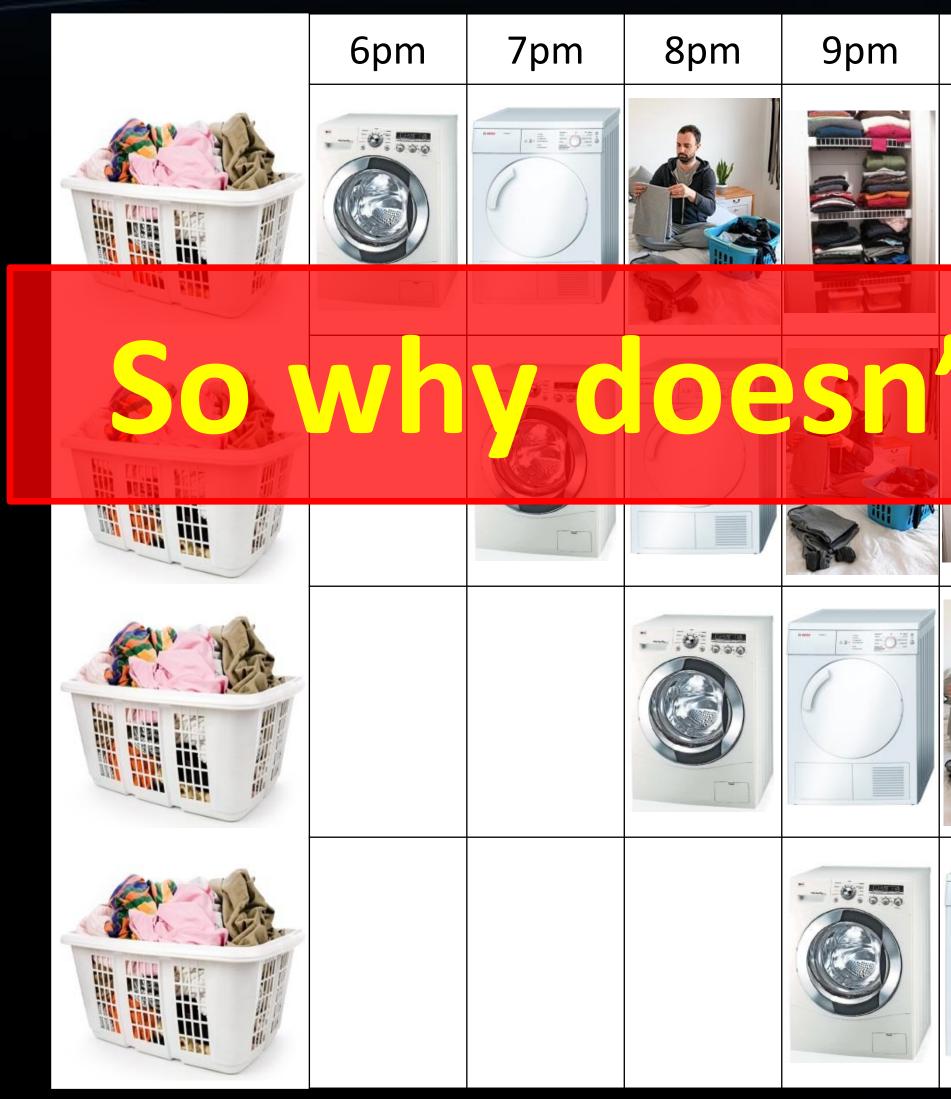
Pipelined processing



10pm	11pm	12pm	01am	02am	03am



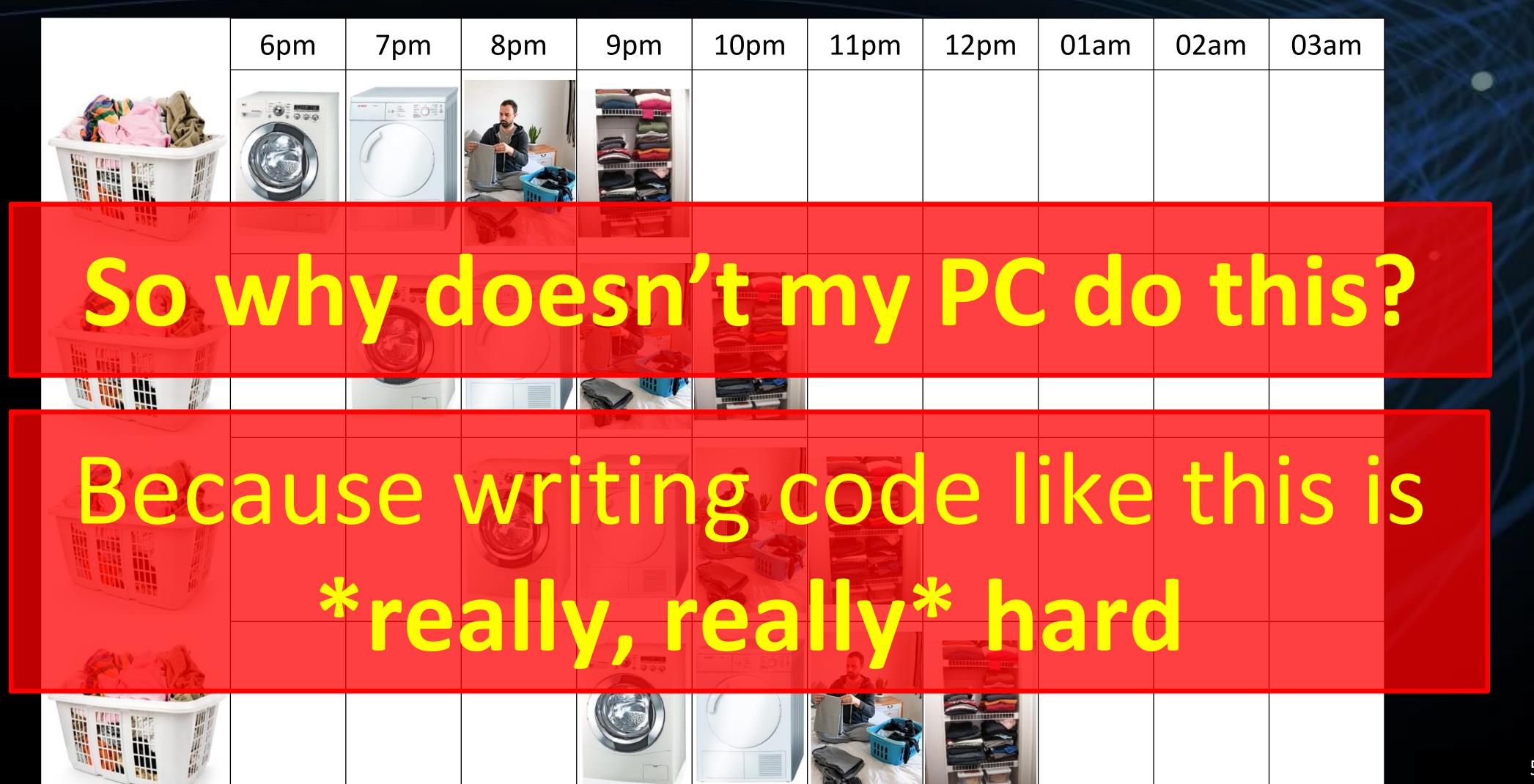
Pipelined processing



10pm	11pm	12pm	01am	02am	03am	
					nis	
						49



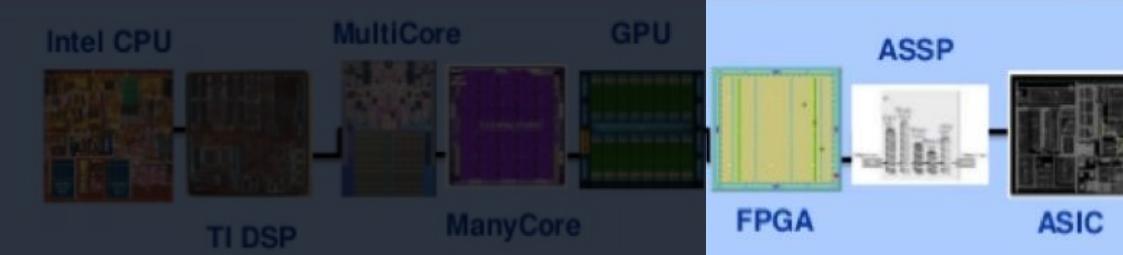
Pipelined processing





What technology can we use?

- Application-specific integrated circuits (ASICs): design encoded into silicon
- Field-programmable gate arrays (FPGAs)



- 69

0

Flexibility, Programming Abstraction

Performance, Area and Power Efficiency

CPU:

- Market-agnostic
- Accessible to many
- programmers (C++)
- Flexible, portable

FPGA:

- Somewhat Restricted Market
- Harder to Program (Verilog)
- More efficient than SW
- More expensive than ASIC

ASIC

- Market-specific
- Fewer programmers
- Rigid, less programmable
- Hard to build (physical)



Programmable circuit-on-chip







- Programmable circuit-on-chip
- Upwards of 9 million logic cells & up to 12000 "math" cells
 - All clocked at ~500MHz
- Up to O(10¹⁵) operations/second
- Fully parallel and fully pipelineable







- Programmable circuit-on-chip
- Upwards of 9 million logic cells & up to 12000 "math" cells
 - All clocked at ~500MHz
- Up to O(10¹⁵) operations/second
- Fully parallel and fully pipelineable









Inen You

IEW data

- Programmable circuit-on-chip
- Upwards of 9 million logic cells & up to 12000 "math" cells
 - All clocked at ~500MHz
- Up to O(10¹⁵) operations/second
- Fully parallel and fully pipelineable







when you

lew data

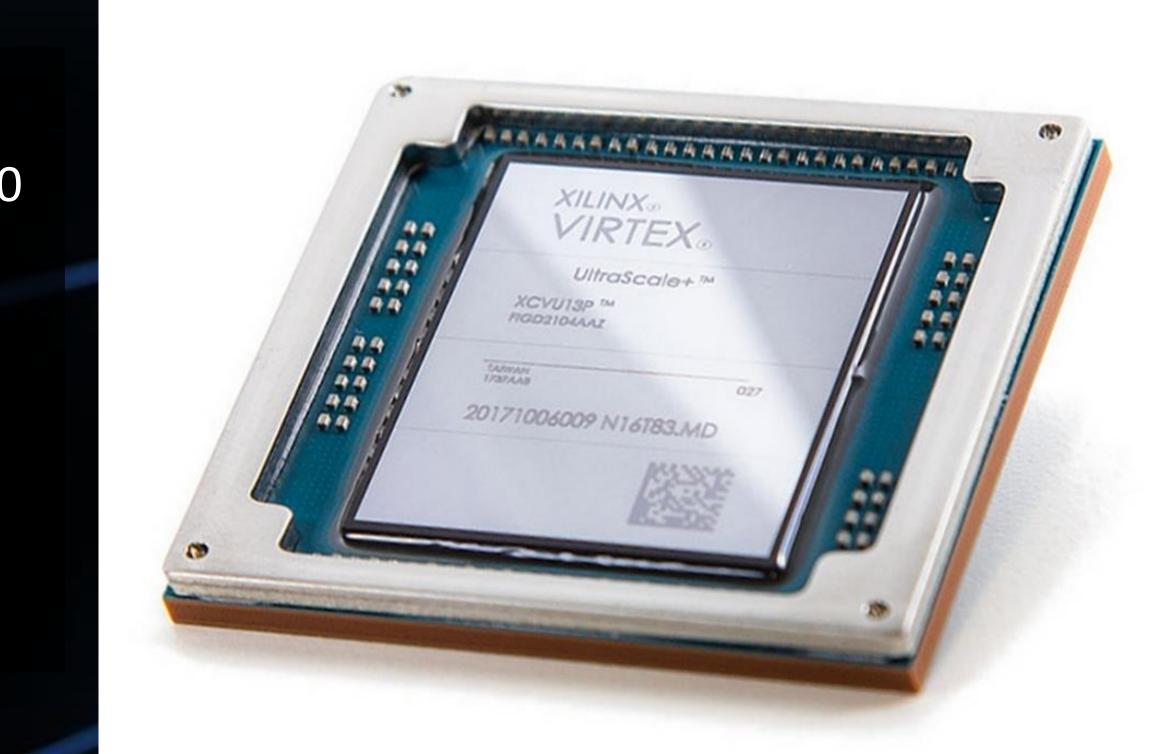
15NS

- Programmable circuit-on-chip
- Upwards of 9 million logic cells & up to 12000 "math" cells
 - All clocked at ~500MHz
- Up to O(10¹⁵) operations/second
- Fully parallel and fully pipelineable

and *really, really* hard to program efficiently

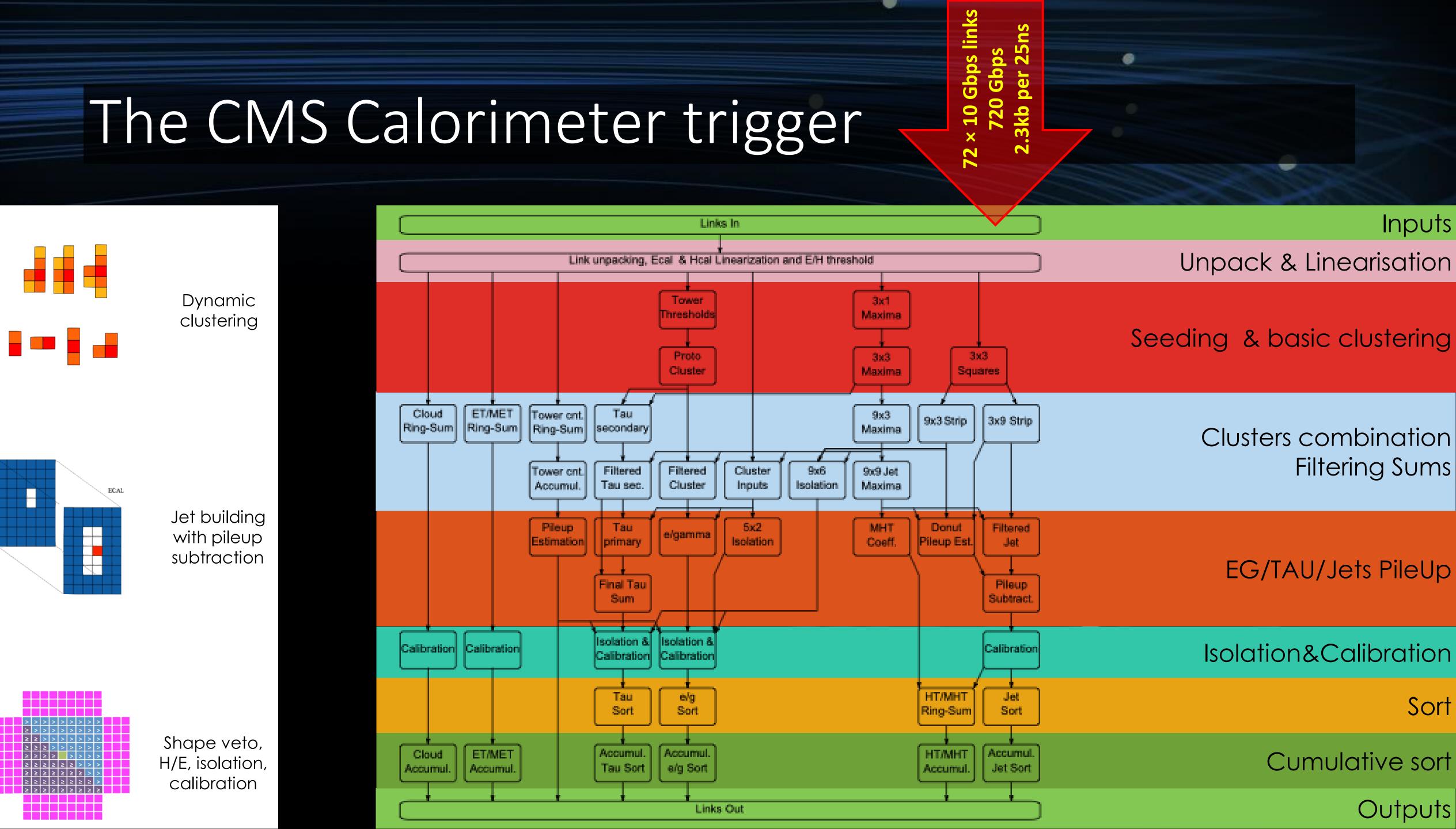
Rather use

dre dealling

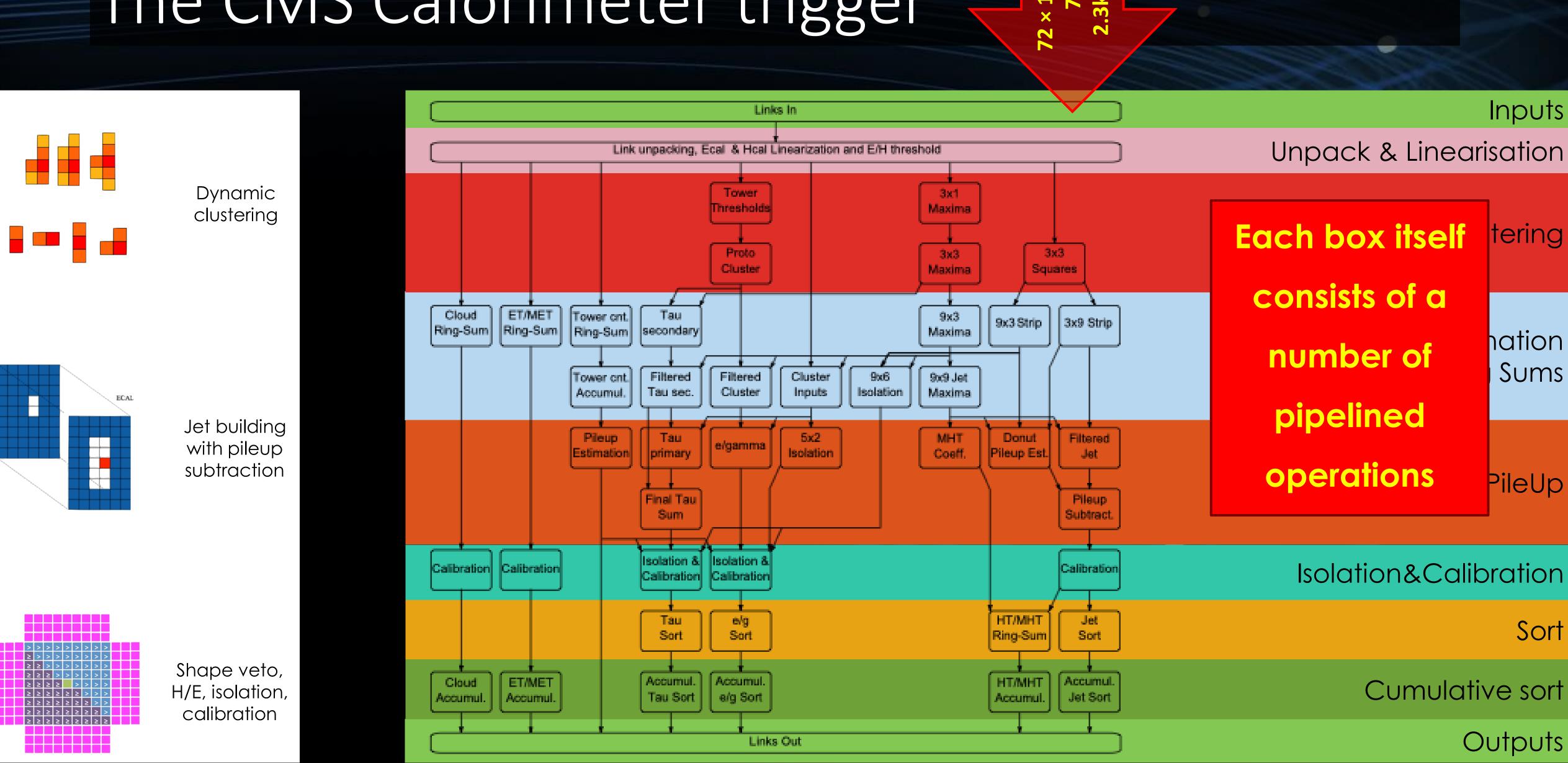








The CMS Calorimeter trigger



72 × 10 Gbps links 720 Gbps 3kb per 25ns 2.3kb

What type of algorithms are we talking about?

- Classical algorithms
 - Clustering in 2D or 3D
 - Pattern identification/matching
 - Kalman filtering

Remember that they have to be implemented in a **fully parallel** and **fully pipelined** form



What type of algorithms are we talking about?

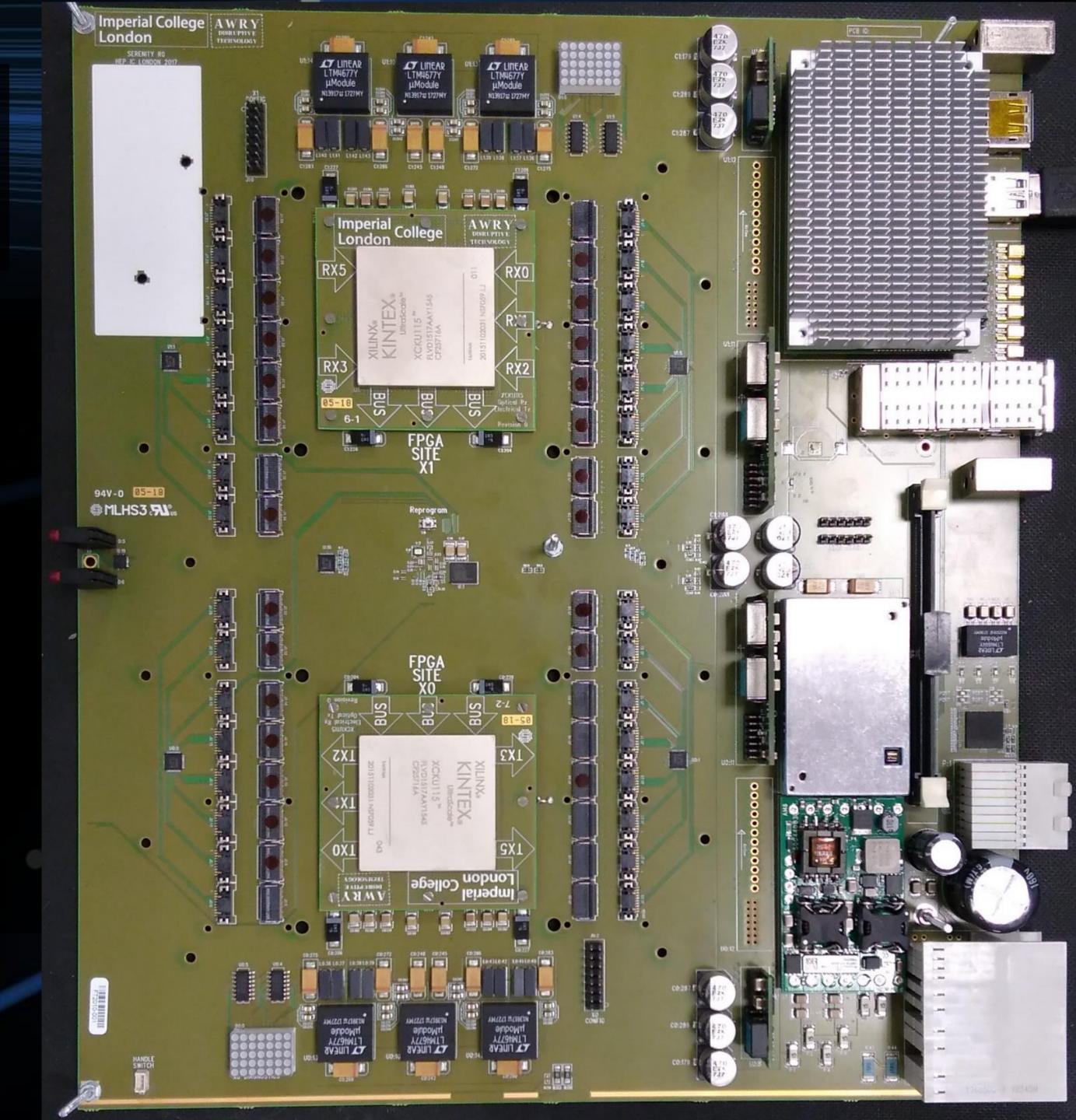
- Classical algorithms
 - Clustering in 2D or 3D
 - Pattern identification/matching
 - Kalman filtering
- In future
 - Particle-flow full reconstruction of particles and events
 - Machine-learning, BDTs and neural-nets in chip

icles and events in chip

Remember that they have to be implemented in a **fully parallel** and **fully pipelined** form



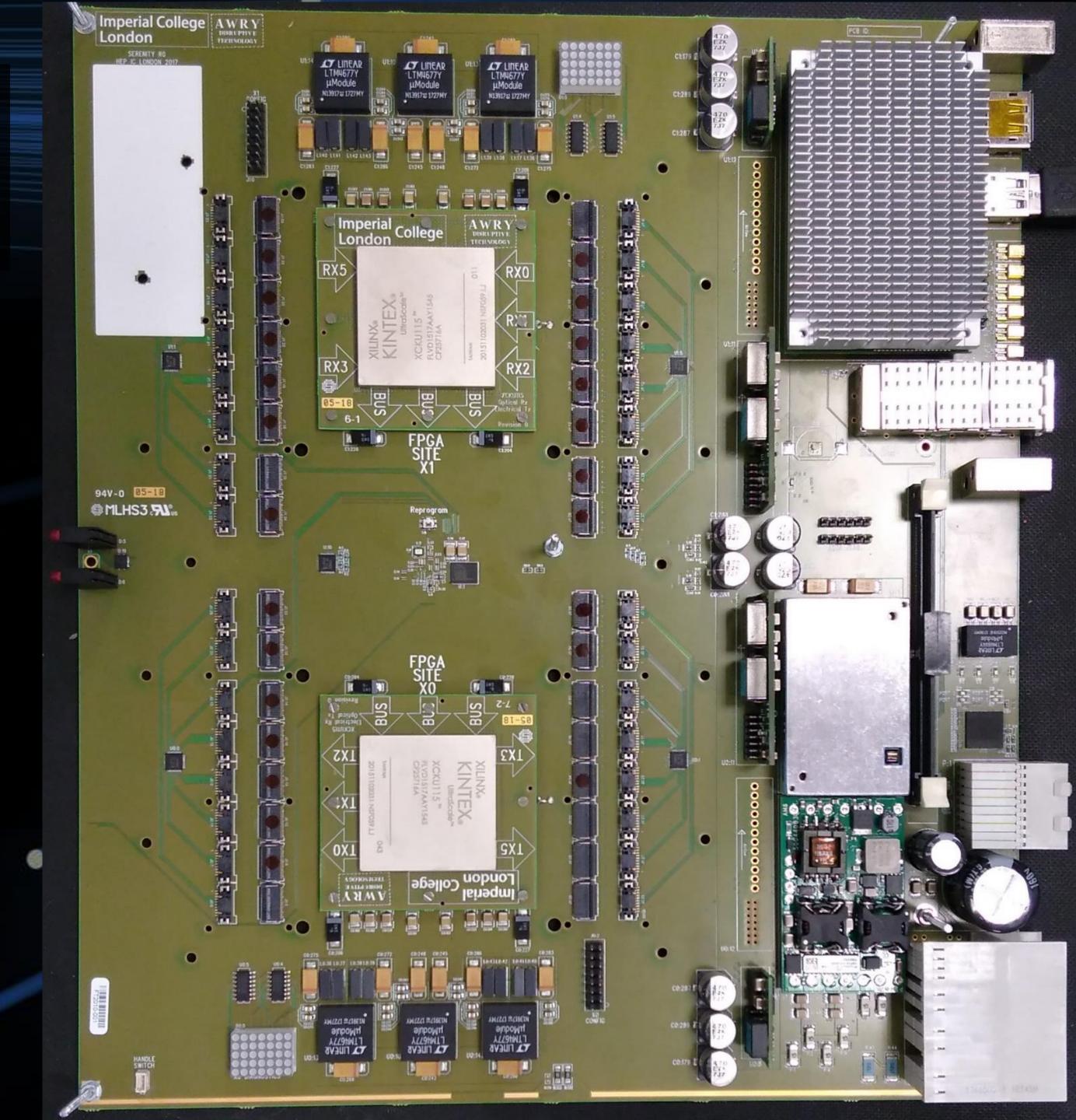
- A chip needs to be attached to something!
- Imperial's latest platform: Serenity



- A chip needs to be attached to something!
- Imperial's latest platform: Serenity
 - 208 optical transmit links &
 208 optical receive links @ 28.5Gbps =

5.9 + 5.9 Tbps

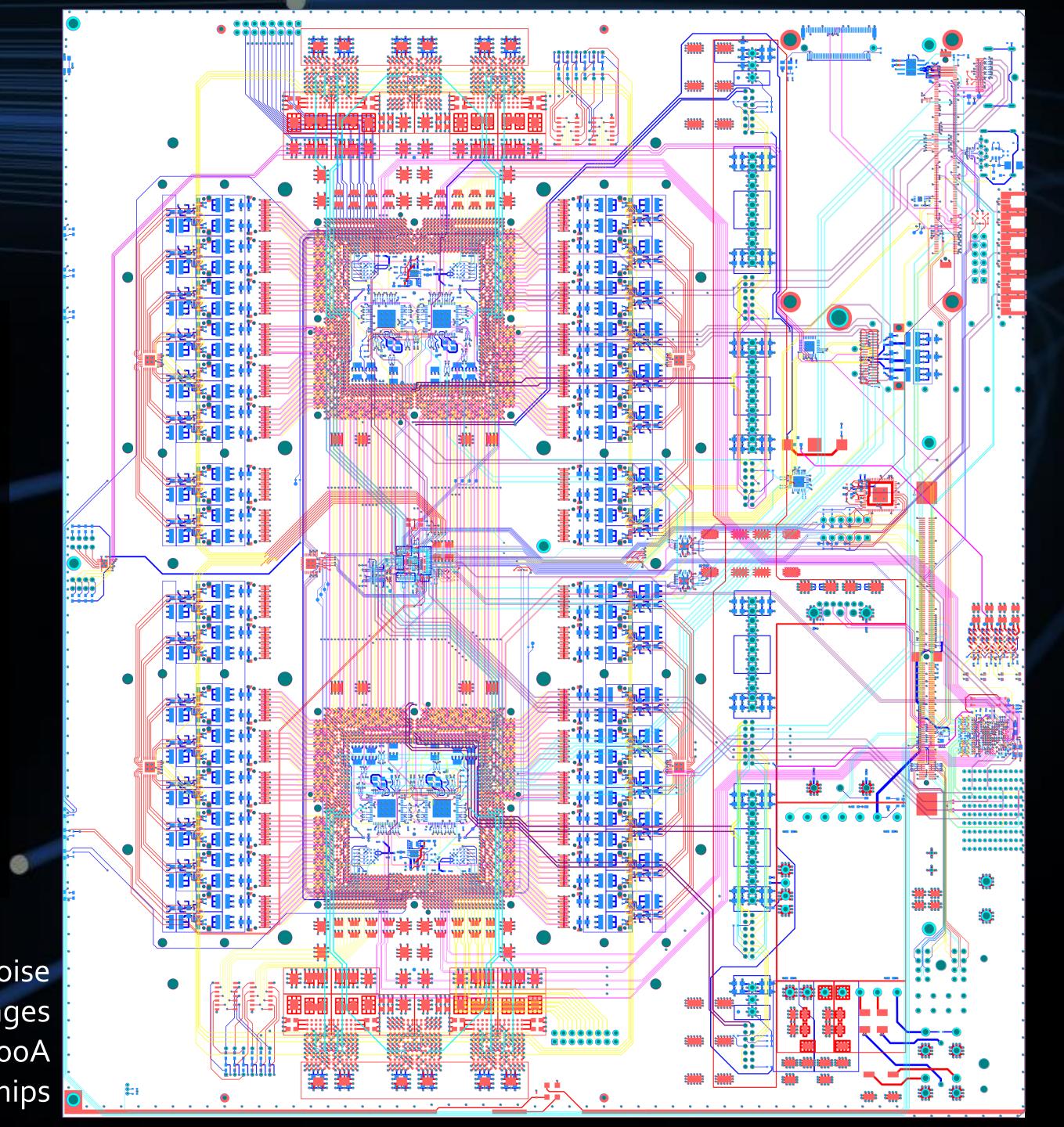
• Skewable to



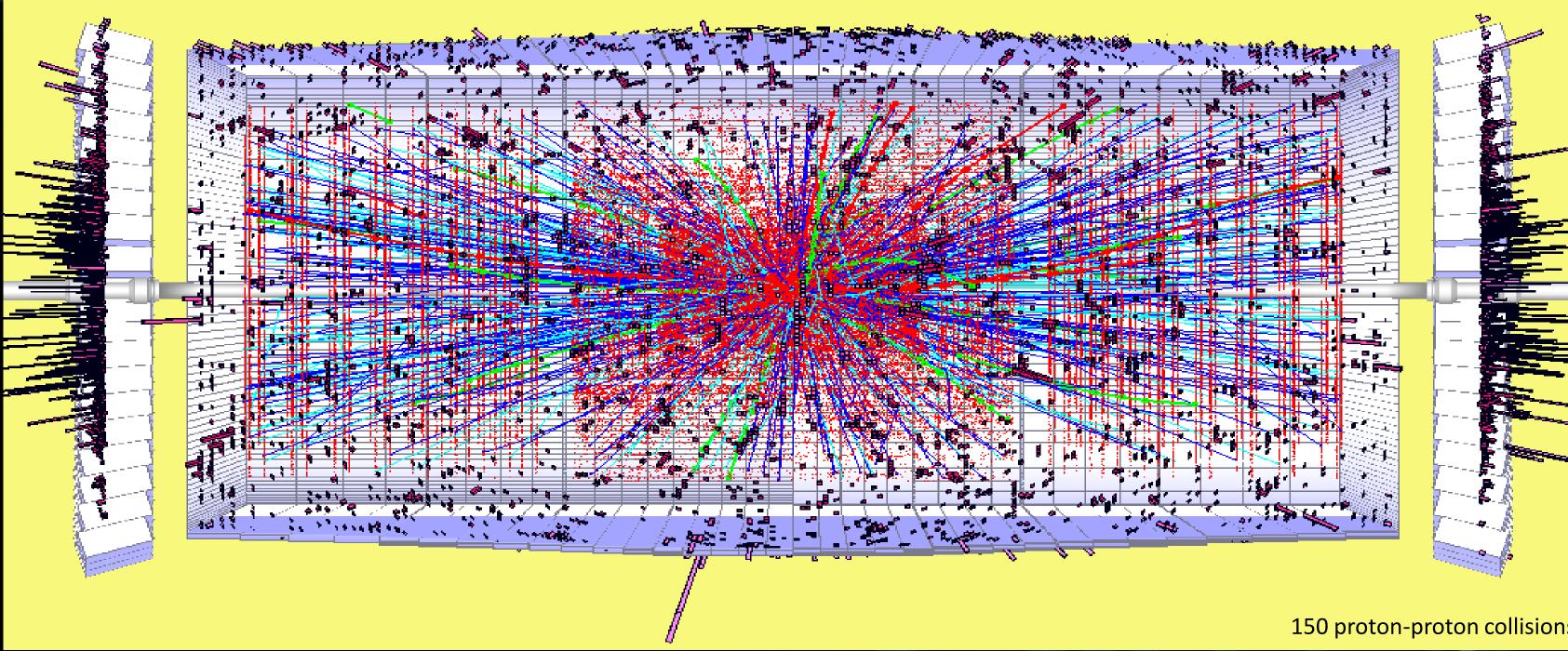
- A chip needs to be attached to something!
- Imperial's latest platform: Serenity

A real challenge:

- 30Gbps links requires~analogue design
- Each chip drawing over 100A of core power
- Heat management a huge challenge
 - 6 ground planes need to control noise
 - 4 power planes need to supply 9 voltages
 - per chip at up to 100A
 - 6 signal layers these are big chips



The future: High-luminosity LHC



- Up to 250 proton-proton collisions per event
- Really, really hard to find the interesting event under all that rubbish
- Need to sift through 300Tbps to find the interesting events

150 proton-proton collisions



To find the interesting events at the LHC we must search through vast amounts of data



- To find the interesting events at the LHC we must search through vast amounts of data
- To do this requires work at the cutting edge of electronics, programming and physics

nust search through vast amounts of data of electronics, programming and physics



- To find the interesting events at the LHC we must search through vast amounts of data
- To do this requires work at the cutting edge of electronics, programming and physics
 - And in the future, also Machine Learning



- To find the interesting events at the LHC we must search through vast amounts of data
- To do this requires work at the cutting edge of electronics, programming and physics
 - And in the future, also Machine Learning
- And this will be even more challenging at the HL-LHC
 - Lots of fun to be had

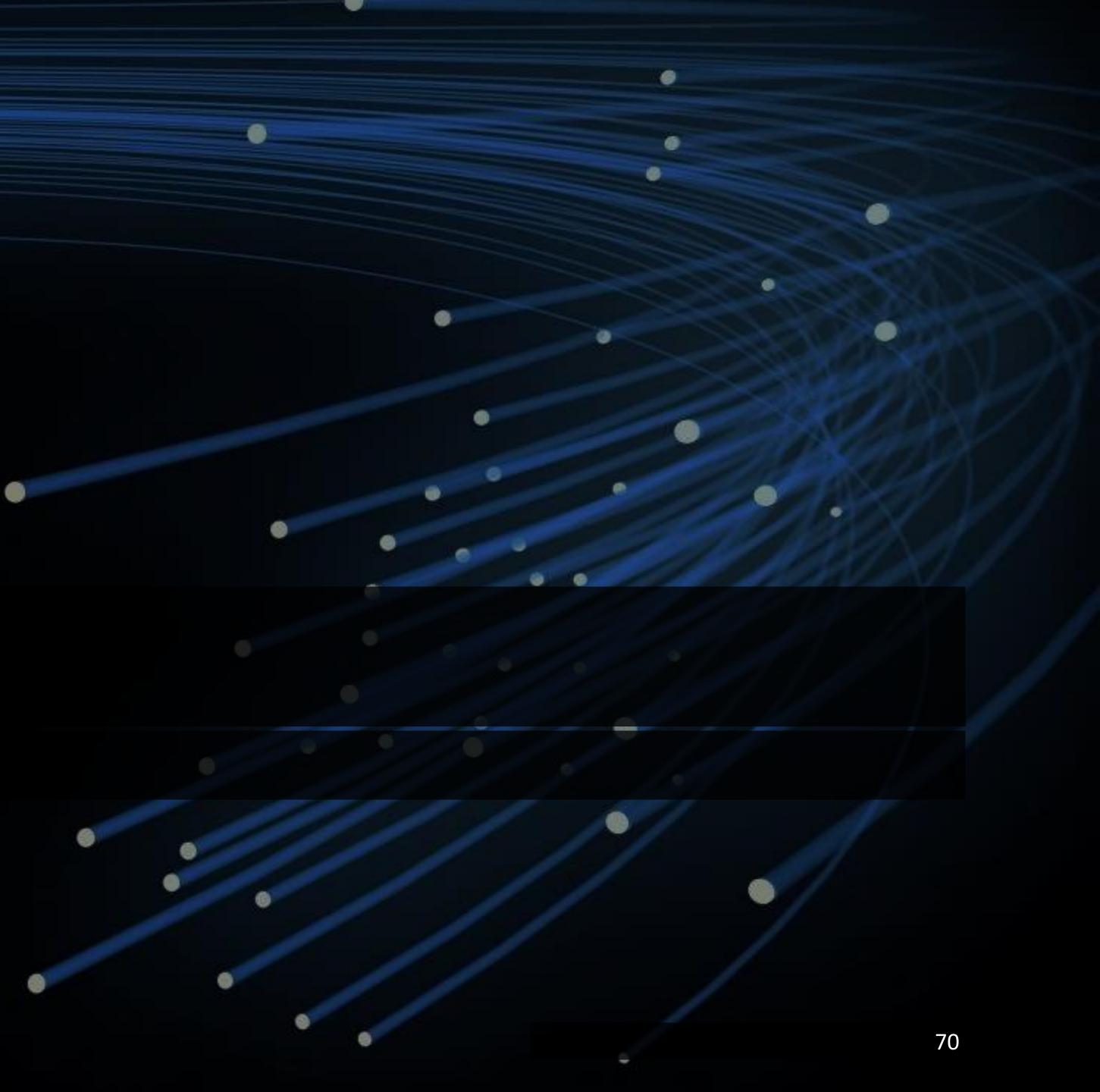


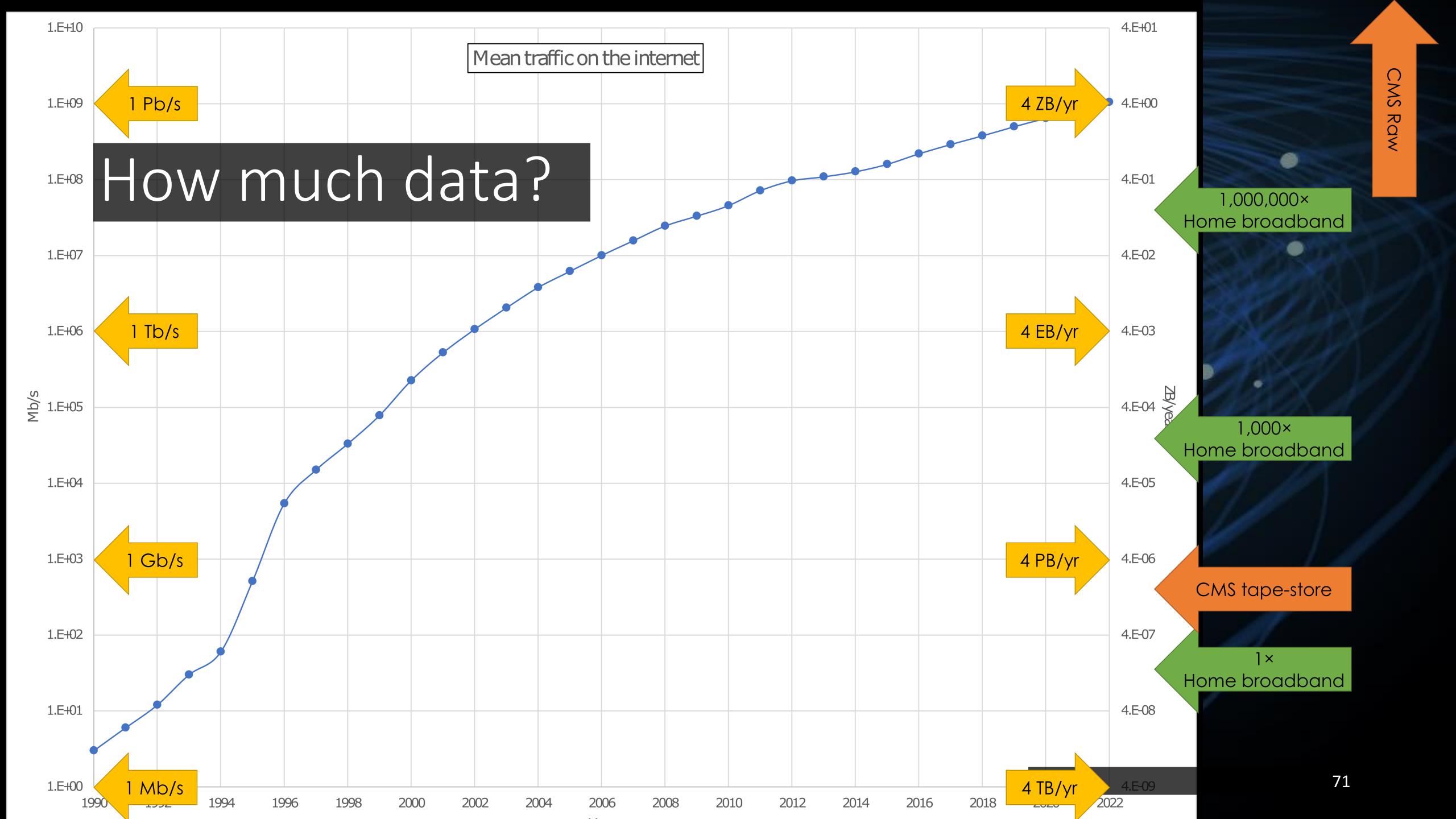
Thanks for listening

Any questions?



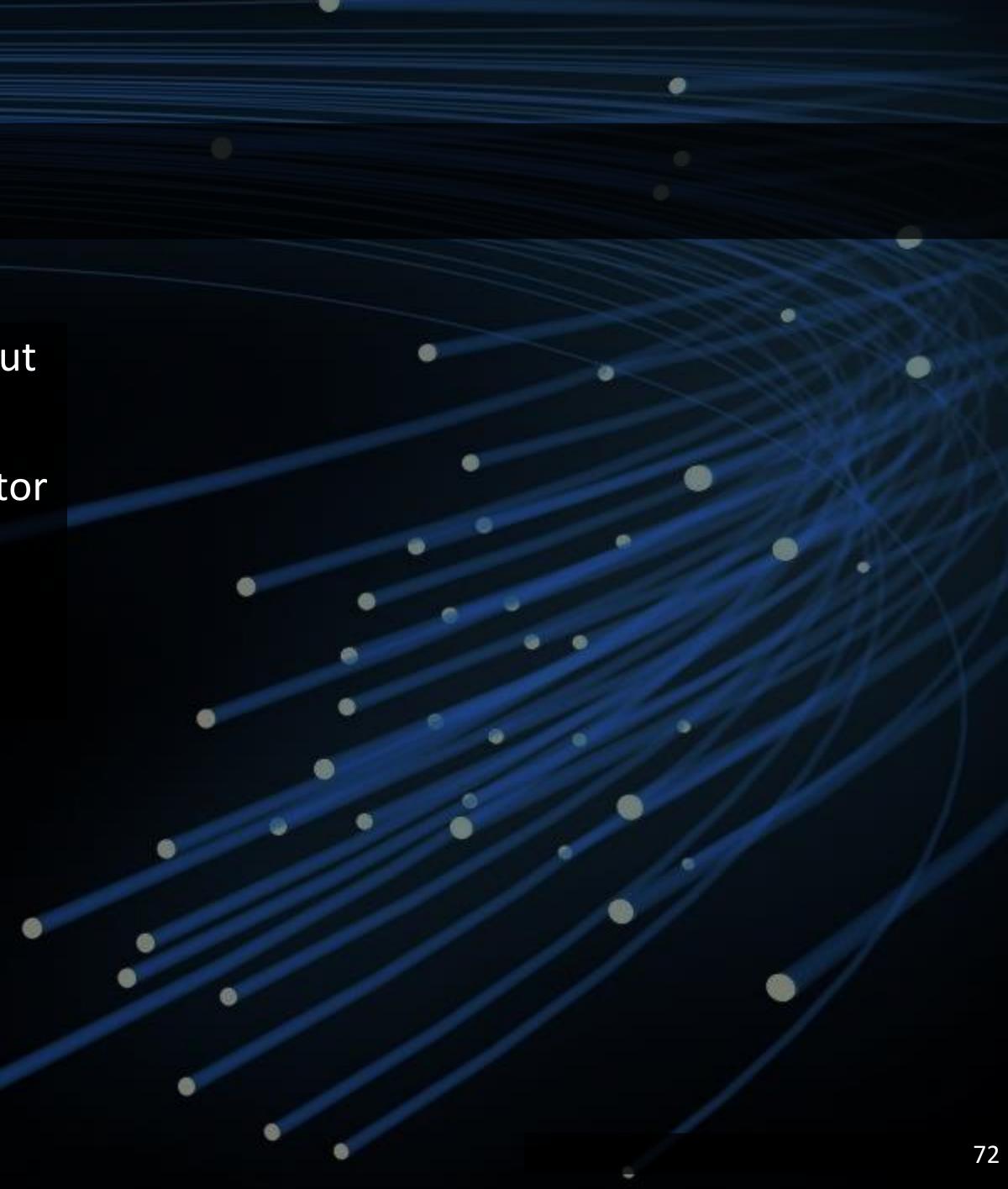
Spares





Layered triggers

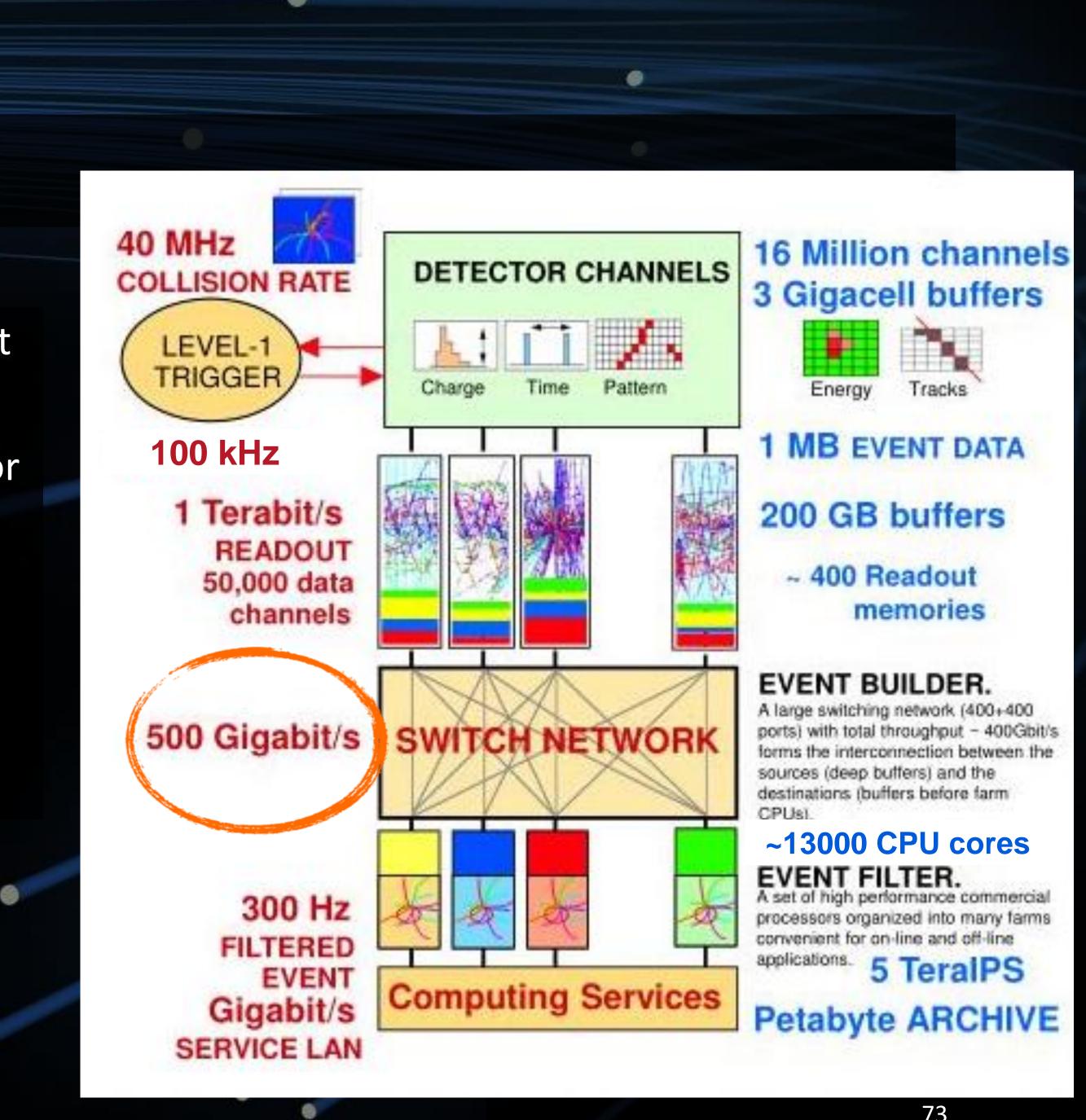
- In our FPGAs we accept events at 100kHz (out of the 40MHz)
 - Reduced the total data volume by a factor of 400
- Small enough to get through an ethernet network into PCs





Layered triggers

- In our FPGAs we accept events at 100kHz (out of the 40MHz)
 - Reduced the total data volume by a factor of 400
- Small enough to get through an ethernet network into PCs
 - Although, of course, "small" here is relative



Spares: Introduction to detectors



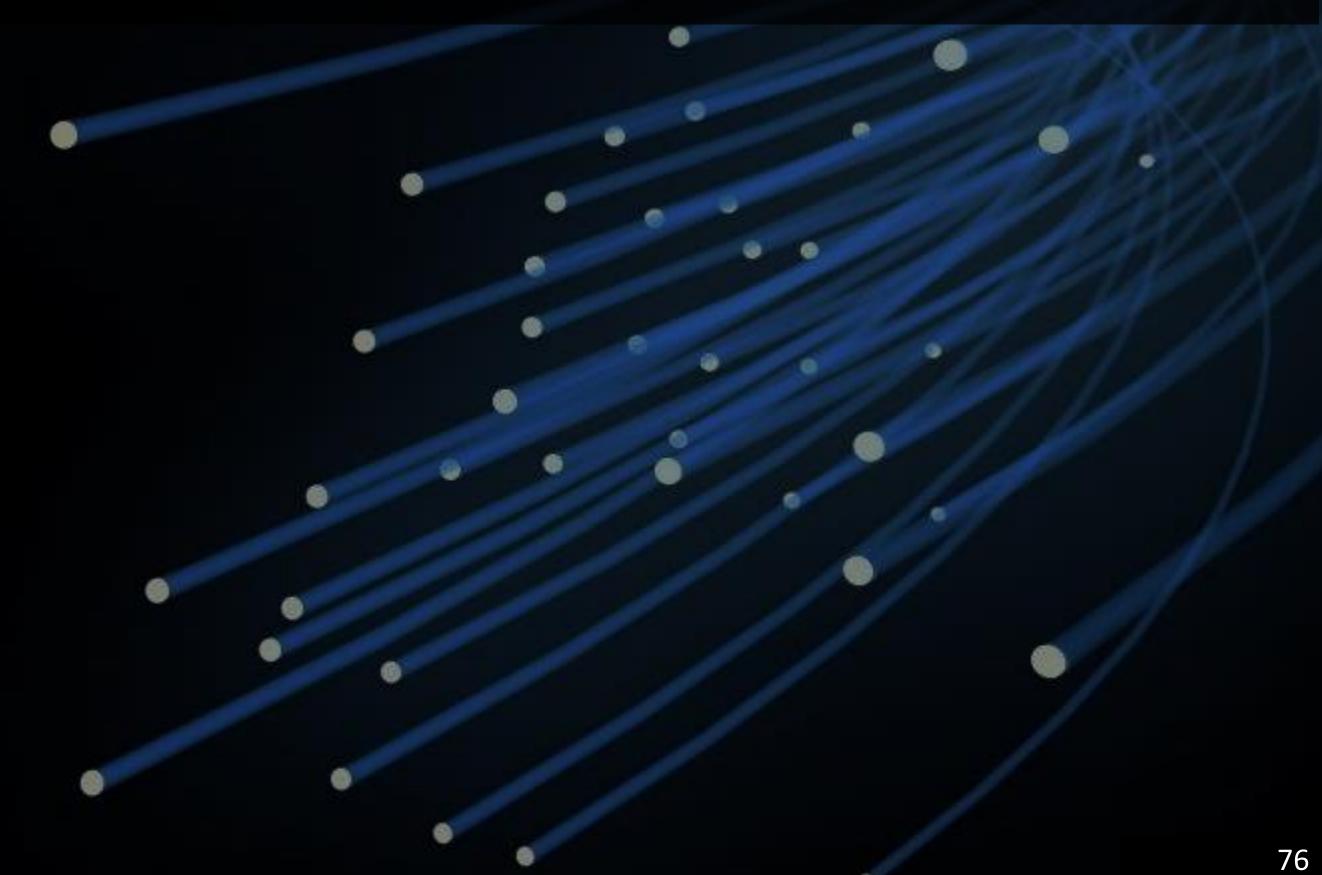
SEE A HIGGS BOSON

ONE DOES NOT SIMPLY



Heavy and unstable

If a heavy and unstable state is produced by a stable particles



• If a heavy and unstable state is produced by a proton-proton collision, it decays quickly into more



Heavy and unstable

- If a heavy and unstable state is produced by a stable particles
- And I mean **REALLY** quickly:

Higgs Boson W/Z Boson Top Quark

Tau Lepton

• Doesn't get anywhere near a detector

• If a heavy and unstable state is produced by a proton-proton collision, it decays quickly into more

 1.6×10^{-22} seconds 3×10^{-25} seconds 5×10^{-25} seconds 2.9×10^{-13} seconds



So we can't see the Higgs (or most other particles) directly

- But you can't "see" the other particles either (in a conventional sense)
- So how are particle detectors built?



A parallel question

 How can you tell the properties of a car and how fast it is going from the outside?



- Take a couple of snapshots
- Work out how it got between them and how long it took to do so

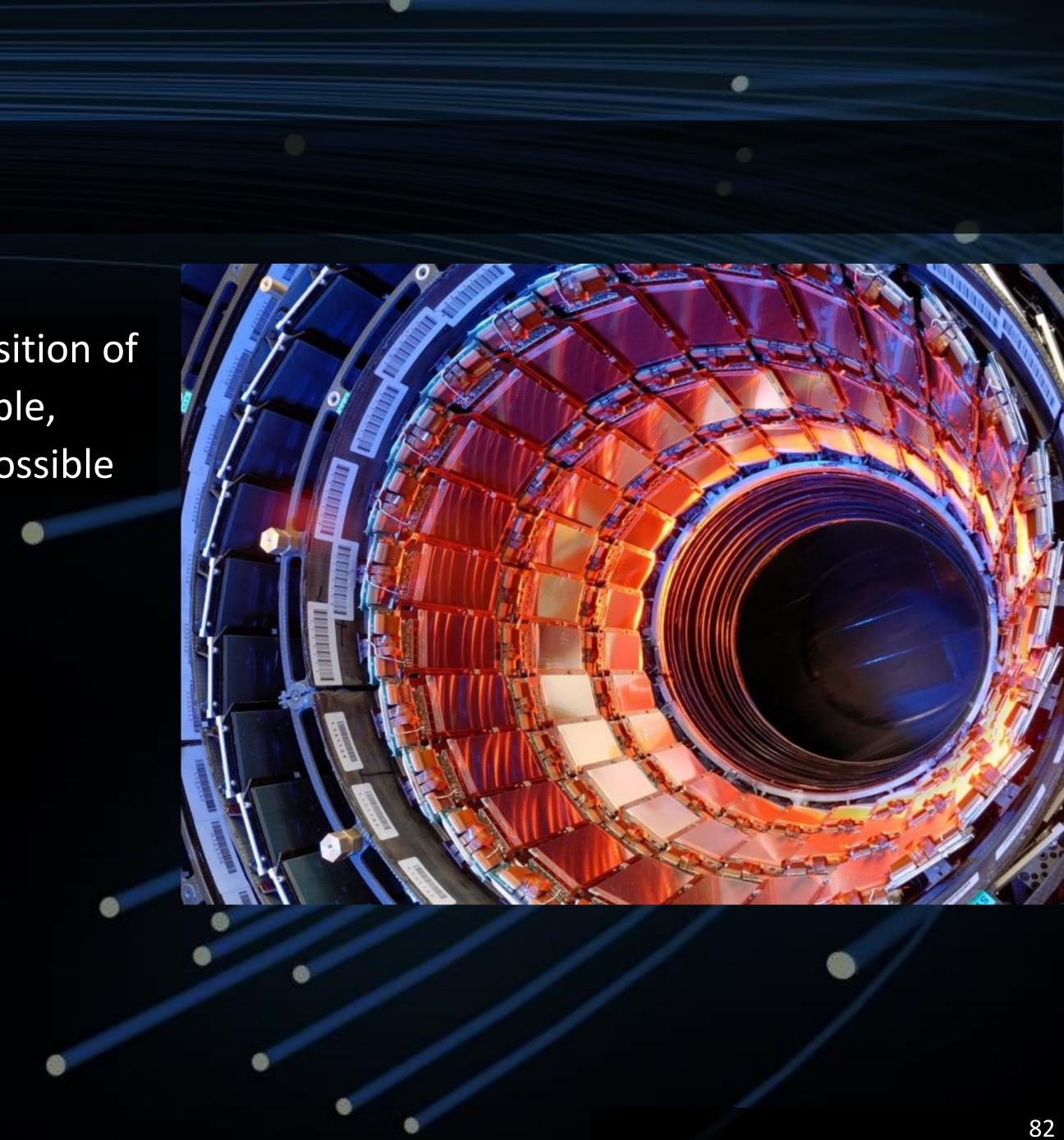


- Take a couple of snapshots
- Work out how it got between them and how long it took to do so
- Does not affect the speed/momentum/energy 0 of the car



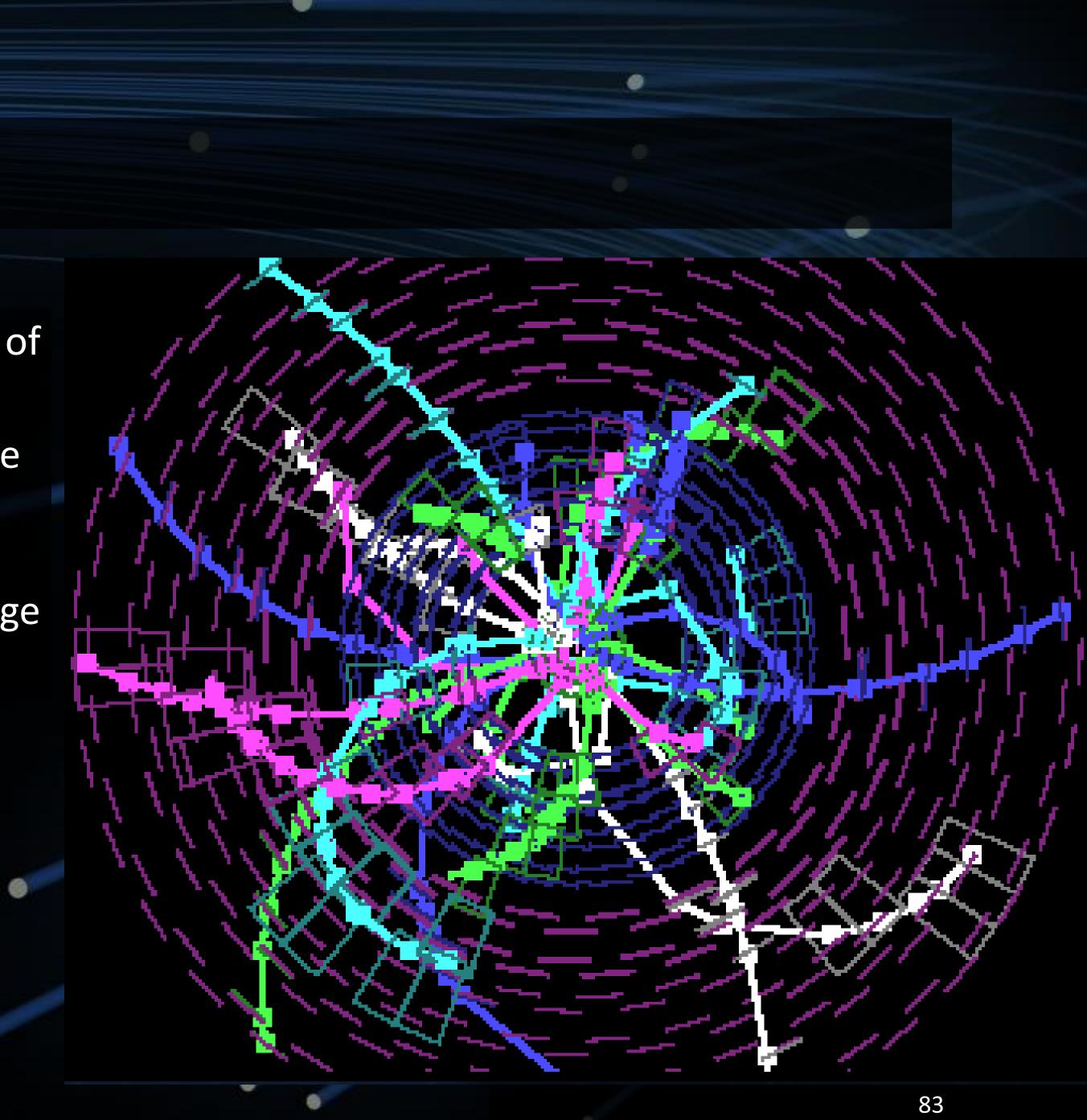


Lightweight tracker that records the position of ulletcharged particles as accurately as possible, while affecting the particle as little as possible





- Lightweight tracker that records the position of charged particles as accurately as possible, while affecting the particle as little as possible
- Join-the-dots
- Apply a magnetic field to determine the charge and momentum



Maximal disruption

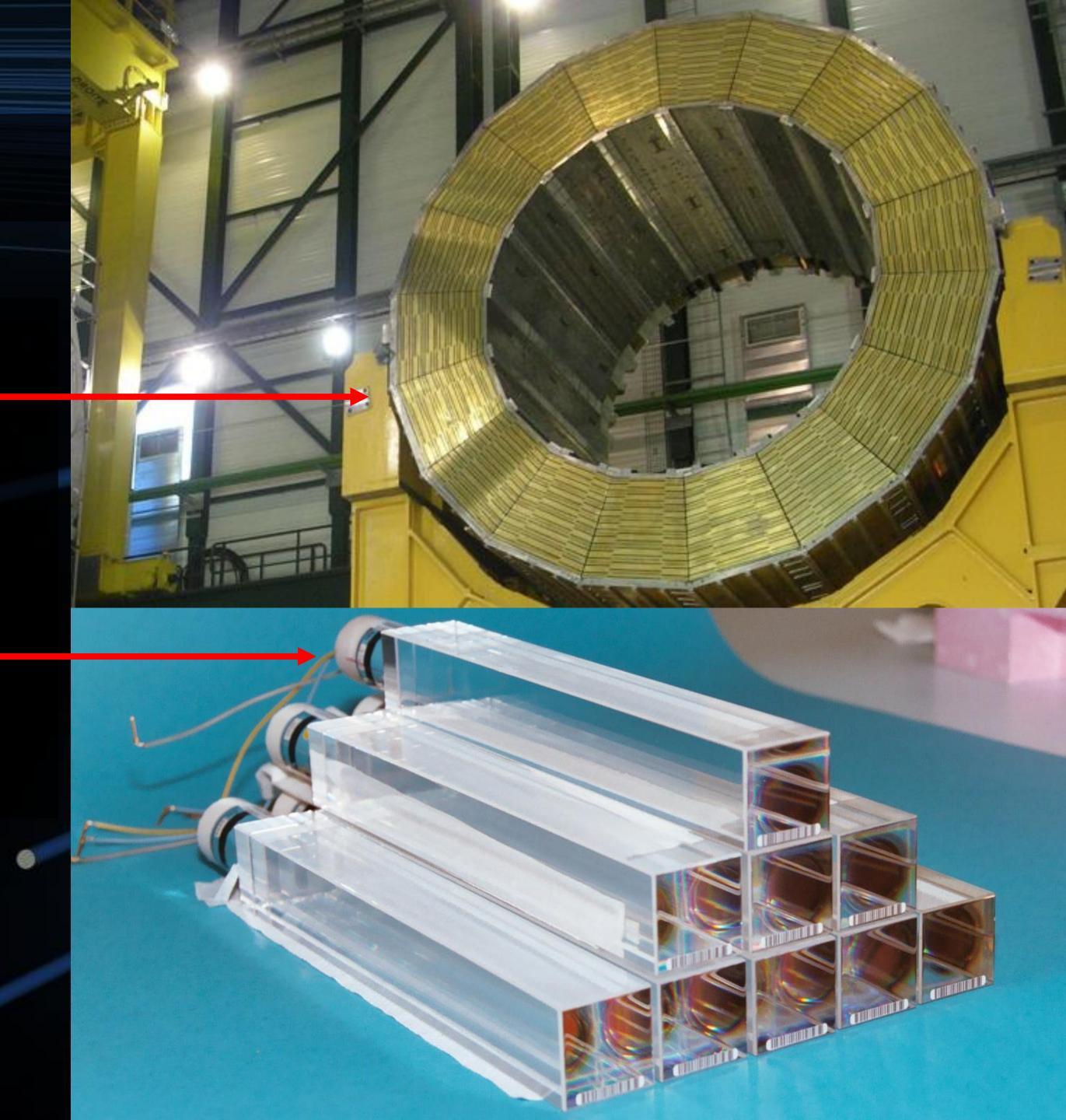
- Place something very heavy in the way
- Collect all the bits; measure how far the pieces get thrown
- Certainly does affect the speed/momentum/ energy of the car





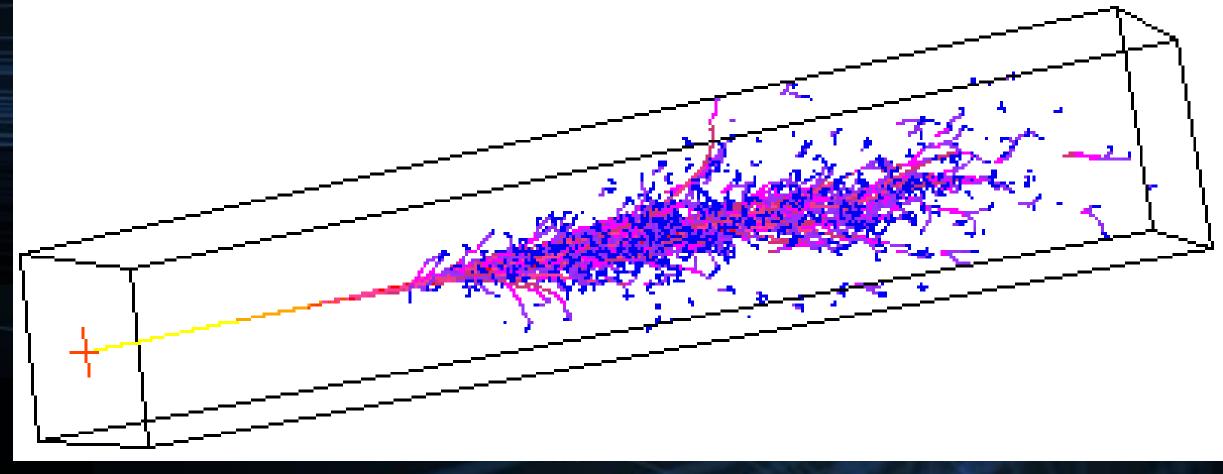
Maximal disruption

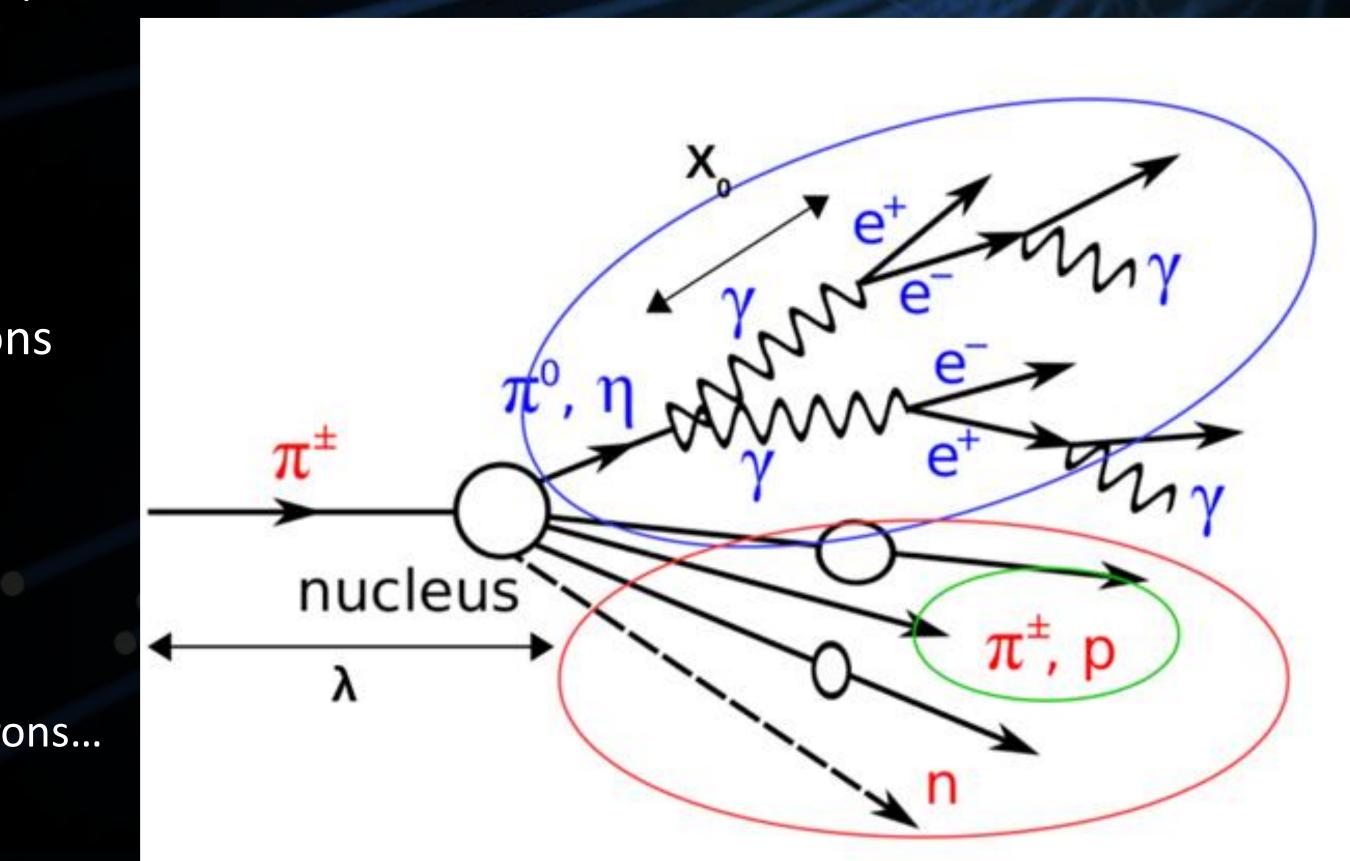
- Put something very dense in the way
 - Brass
 - Steel
 - Tungsten
 - Depleted uranium
 - Lead tungstate crystals
- Catch the light in some transparent medium using photomultipliers



Maximal disruption

- Energetic leptons particles emit photons ullet
 - Energetic photons pair-produce electrons and positrons ullet
 - Which emit photons ullet
 - Which pair-produce electrons and positrons \bullet
 - Which emit photons igodot
 - Which pair-produce electrons and positrons \bullet
 - Which emit photons...
- Energetic hadrons break up into lighter hadrons ightarrow
 - Which break up into lighter hadrons
 - Which break up into lighter hadrons \bullet
 - Which break up into lighter hadrons... \bullet
 - Pions decay to photons
 - Which pair-produce electrons and positrons...
 - Which emit photons...





Particles collide here



Pixel and strip trackers



Calorimeters



4T superconducting magnet

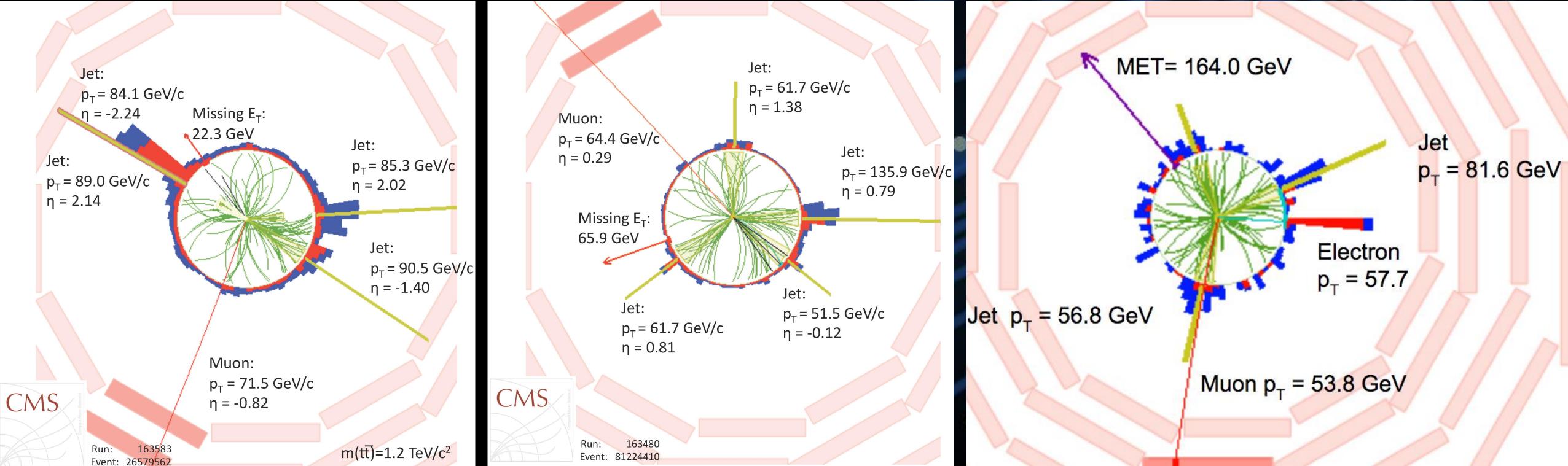


Muon trackers



Event reconstruction

- Join the tracks with the energy deposits
- Apply energy and momentum conservation to reconstruct everything all the way down to the interaction point





Spares: Introduction to triggers

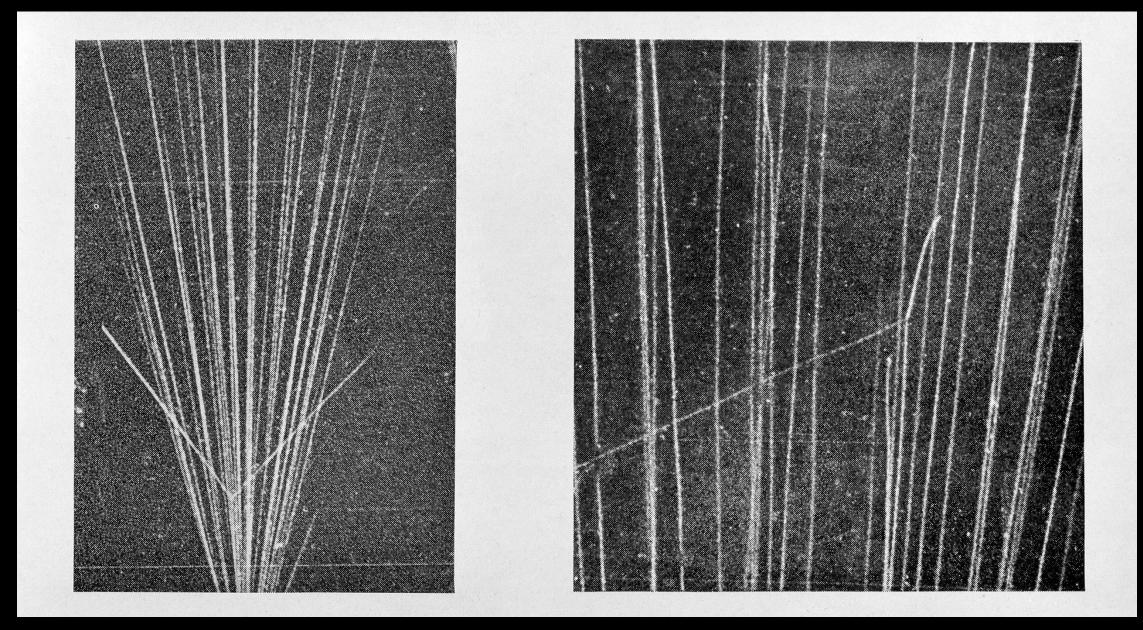


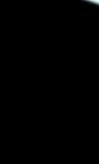
• Trigger basic requirements

- Need high efficiency for selecting processes for physics analysis.
- Need large reduction of rate from unwanted high-rate processes
- Robustness is essential
- Highly flexible, to react to changing conditions
- System must be affordable

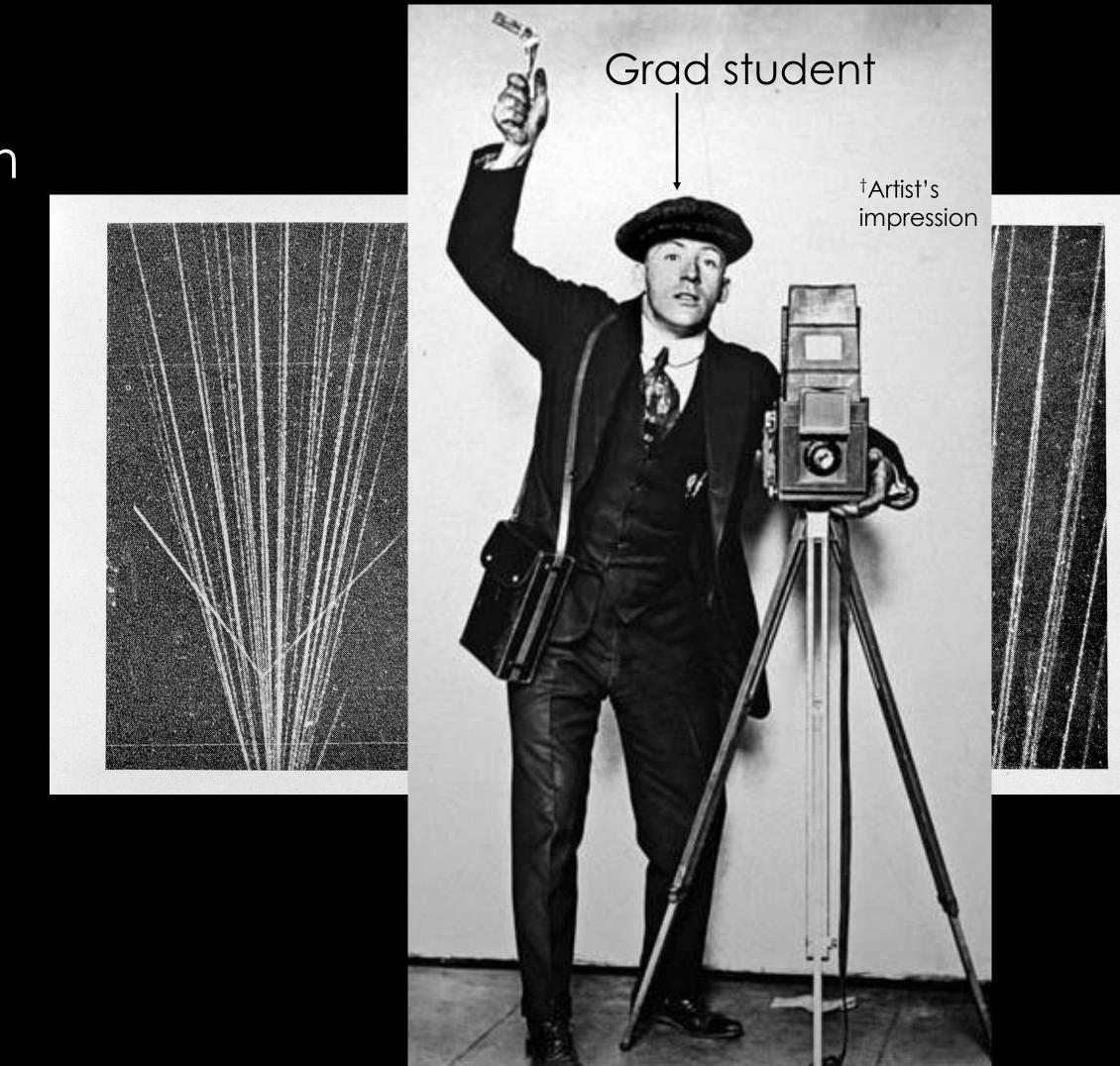
REMINDER

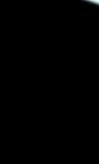
- Cloud-chamber images recorded on film
- Need some way to trigger the camera





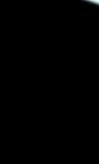
- Cloud-chamber images recorded on film
- Need some way to trigger the camera





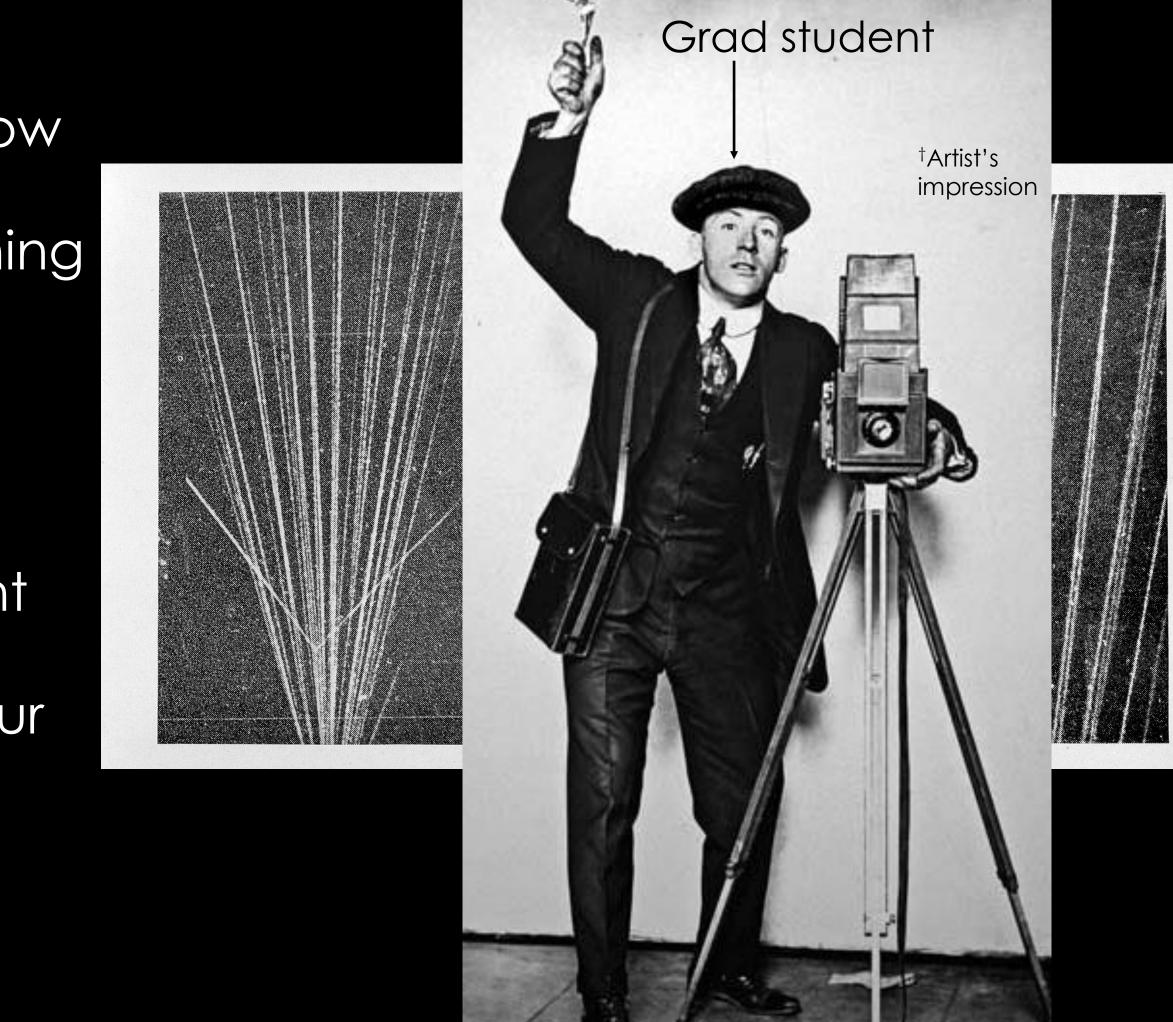
- High efficiency? Nope reflexes too slow
- Large rate reduction? Better than nothing
- Robustness? No keep wanting sleep, coffee, toilet breaks, etc.
- Highly flexible? Depends on the student

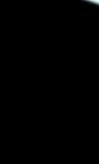




- High efficiency? Nope reflexes too slow
- Large rate reduction? Better than nothing
- Robustness? No keep wanting sleep, coffee, toilet breaks, etc.
- Highly flexible? Depends on the student
- Affordable? Well that's one thing in your favour, I suppose







- High efficiency?
- Large rate reduce
- Robustness? No coffee, toilet bre

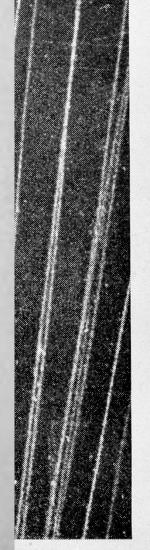
Although Rutherford & Geiger did note that "Strong coffee with a pinch of Strychnine" improved the ability to spot scintillation light

- Highly flexible? Depends on the student
- Affordable? Well that's one thing in your favour, I suppose

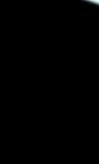
THE EARLIEST TRIGGER

Grad student

[†]Artist's impression







- Blackett pioneered a technique to trigger ightarrowthe camera of cloud chambers (and got the Nobel prize for this and other work)
- Just missed out on discovering the positron in 1932
- Stevenson and Street used this to confirm the discovery of the muon in 1937

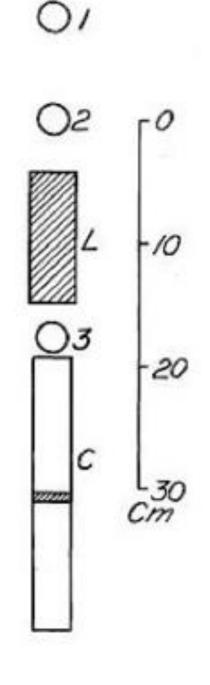
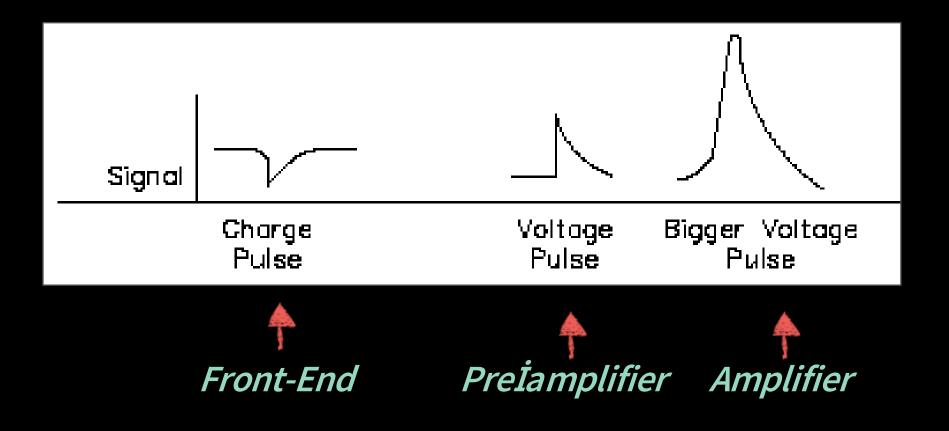




FIG. 1. Geometrical arrangement of apparatus

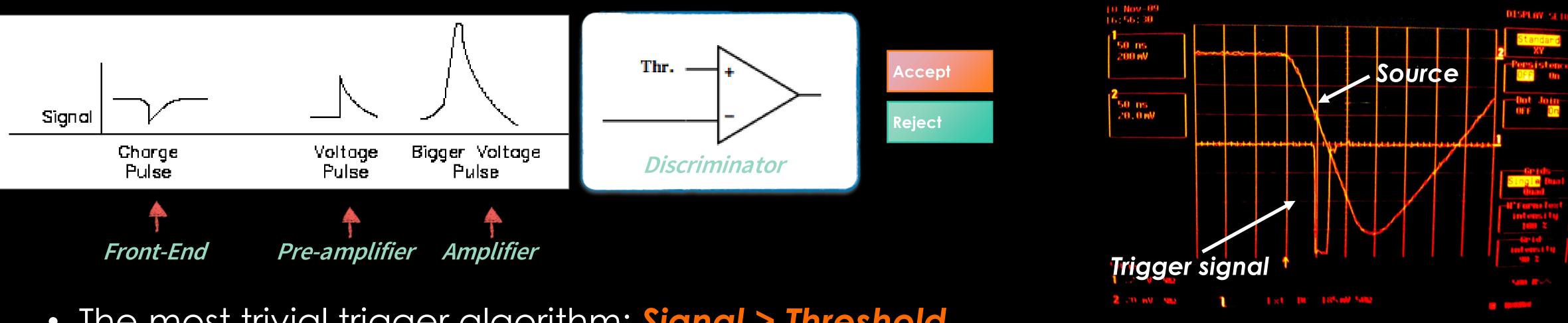
THE SIMPLEST TRIGGER SYSTEMS

- Source: Use the signals from the Front-End of the detectors themselves
 - Binary: tracking detectors (pixels, strips)
 - Analog: tracking detectors, time of flight detectors, calorimeters, ...



THE SIMPLEST TRIGGER SYSTEMS

- Source: Use the signals from the Front-End of the detectors themselves
 - Binary: tracking detectors (pixels, strips)
 - Analog: tracking detectors, time of flight detectors, calorimeters, ...

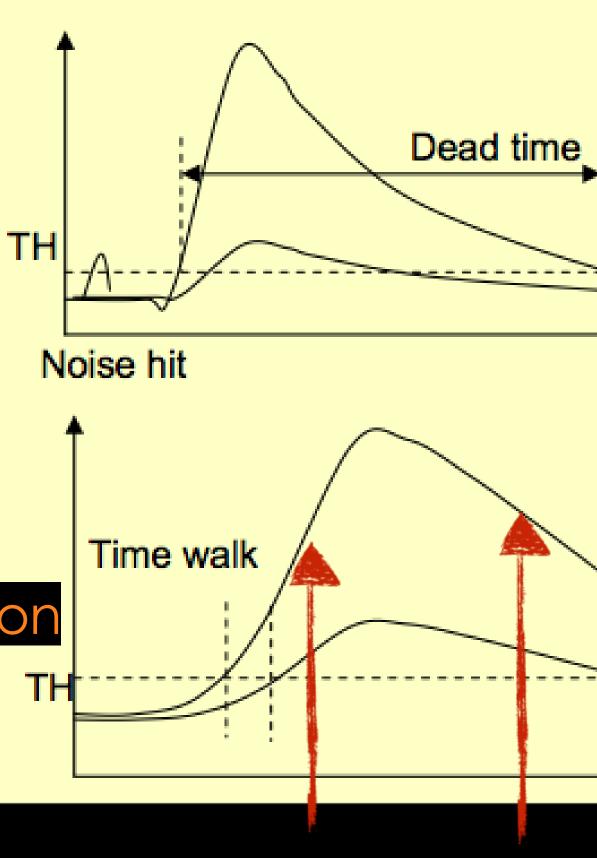


- The most trivial trigger algorithm: Signal > Threshold
 - Apply the lowest possible threshold
 - Identify best compromise between hit efficiency and noise rate \bullet



DETECTOR SIGNALS CHARACTERISTICS

- Pulse width
 - Limits the effective hit rate
 - Must be adapted to the desired trigger rate \bullet
- Time walk
 - The threshold-crossing time depends on the signal amplitude
 - Must be minimal good trigger systems ightarrow
- Time walk can be suppressed by triggering on total signal fraction
 - Applicable on same-shape input signals with different amplitude
 - Scintillator detectors and photomultipliers

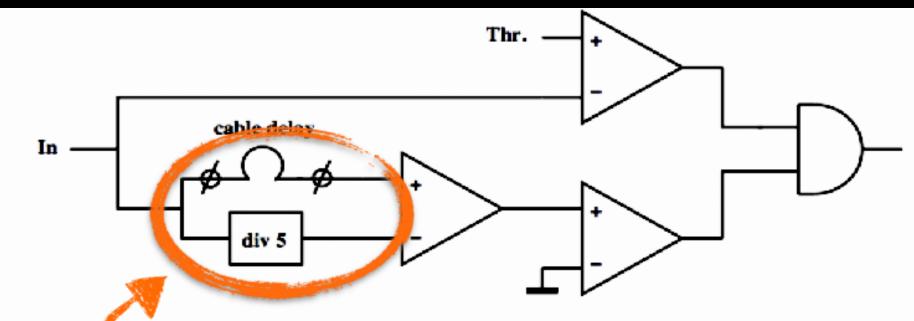


Leading edge

Trailing edge



THE CONSTANT FRACTION DISCRIMINATOR

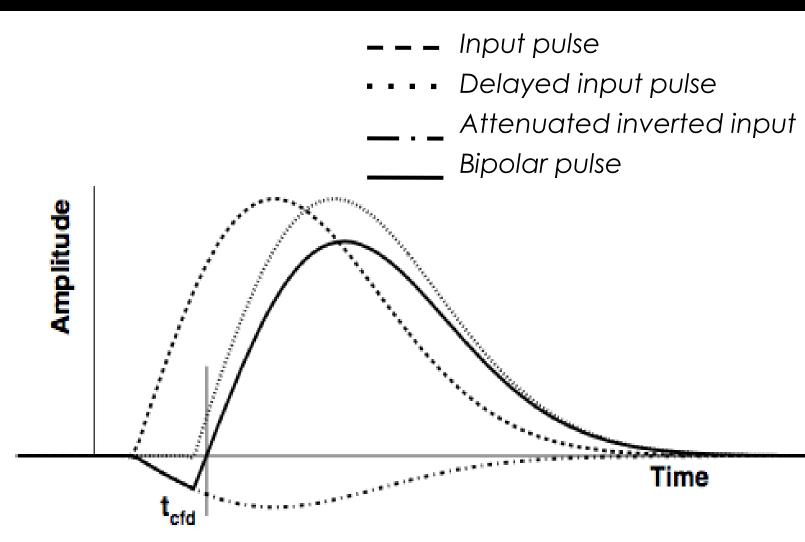


- Attenuation + configurable delay applied before the discrimination determines *t*_{CFD}
- If delay too short, the unit works as a normal discriminator since the output of the normal discriminator fires later than the CFD part

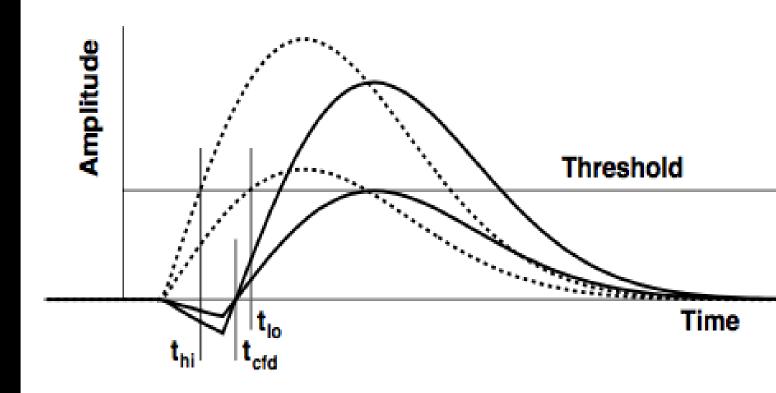
Signals with the same rising time, at a fraction f

 $\frac{A(t)}{f} - A(t - \Delta t) = 0 \quad \text{at } t = t_{\text{cfd}}$

 $\Delta t_f = t(f \cdot A_0) - t(A_0) = \text{const.}$

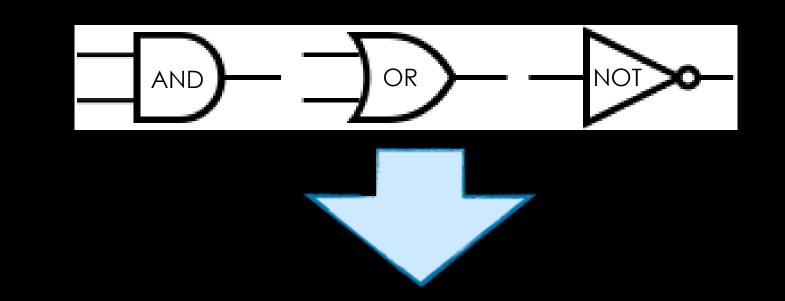


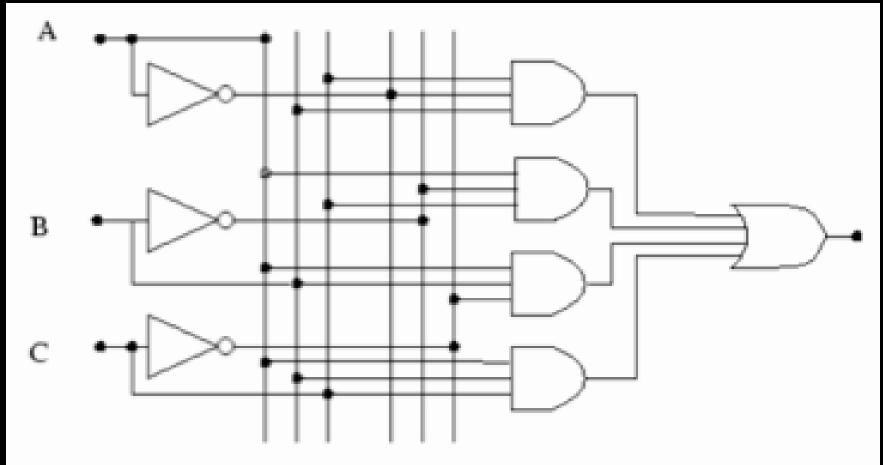
The output of the CFD fires when the bipolar pulse changes polarity



TRIGGER LOGIC IMPLEMENTATION

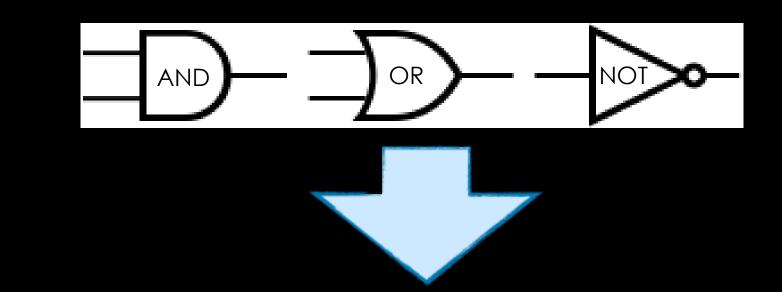
- Once we are in the digital domain, all manipulations can be broken down to a Boolean operations
- Combinatorial
 - Summing, Decoders, Multiplexers,...
- Sequential
 - Flip-flops, Registers, Counters,...



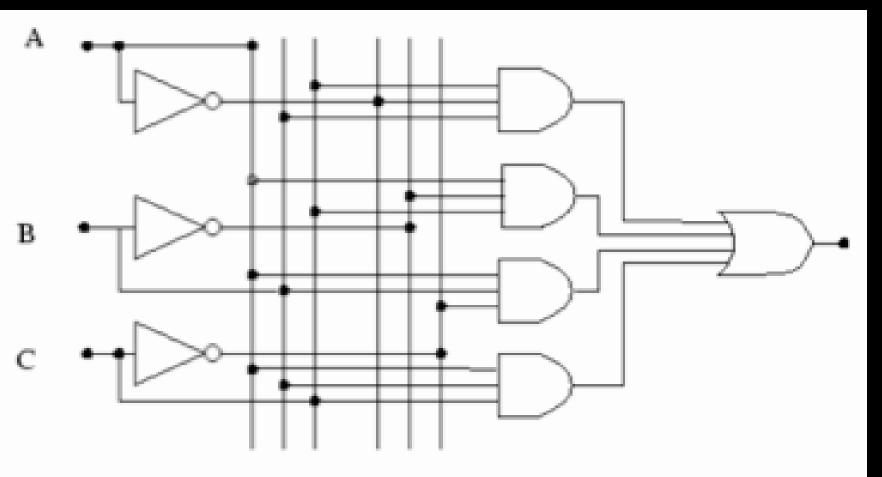


TRIGGER LOGIC IMPLEMENTATION

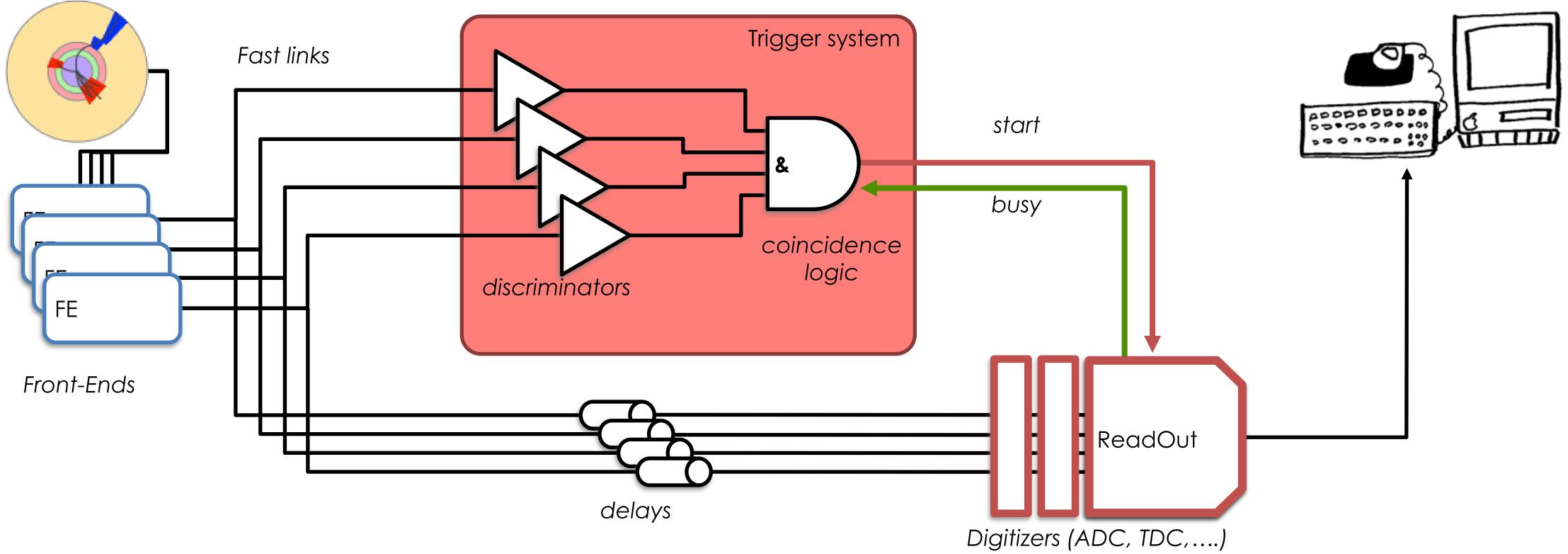
- Once we are in the digital domain, all manipulations can be broken down to a Boolean operations
- Combinatorial
 - Summing, Decoders, Multiplexers,...
- Sequential
 - Flip-flops, Registers, Counters,...



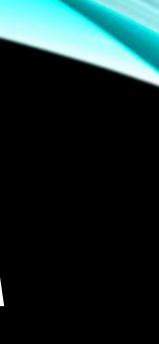
- Data propagates
- as a wave
- through the logic
- Operations
- happen at well
- defined times
- and in a well
- defined order







A SIMPLE TRIGGER SYSTEM





- The key parameter in high speed trigger systems design
 - The fraction of the acquisition time when no events can be recorded.
 - Typically of the order of few %
 - Reduces the overall system efficiency
- Arises when a given processing step takes a finite amount of time
 - Readout dead-time
 - Trigger dead-time
 - Operational dead-time

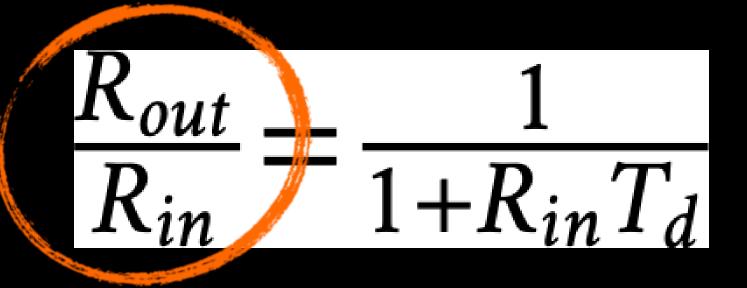
DEADTIME

- Writing to disk or tape is much slow than accepting data into RAM
- If you select an event and start writing it to disk, you cannot accept any more events until you finish writing, even if they are interesting

DEADTIME EXAMPLE

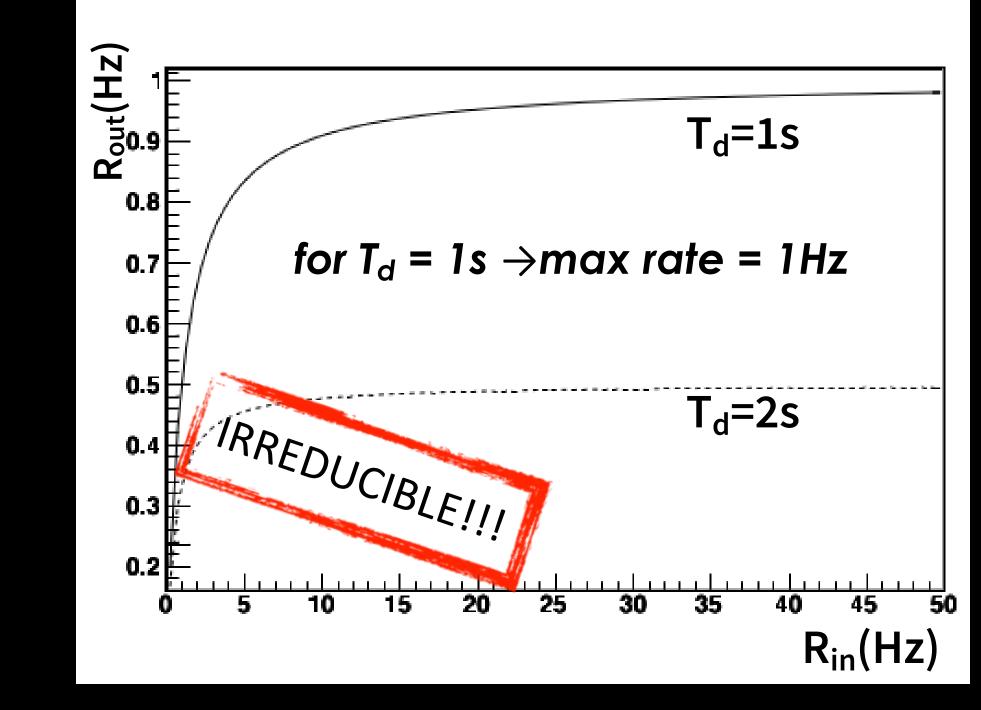
- For input rate, " R_{in} ", Readout rate, " R_{out} ", and time taken to write to disk, " T_d "
- Fraction of lost events = $R_{out} \cdot T_d$
- Event output rate $R_{out} = (1-R_{out} \bullet T_d) \bullet R_{in}$

Fraction of surviving events



To achieve high efficiency $\Rightarrow R_{in} \cdot T_d \ll 1$

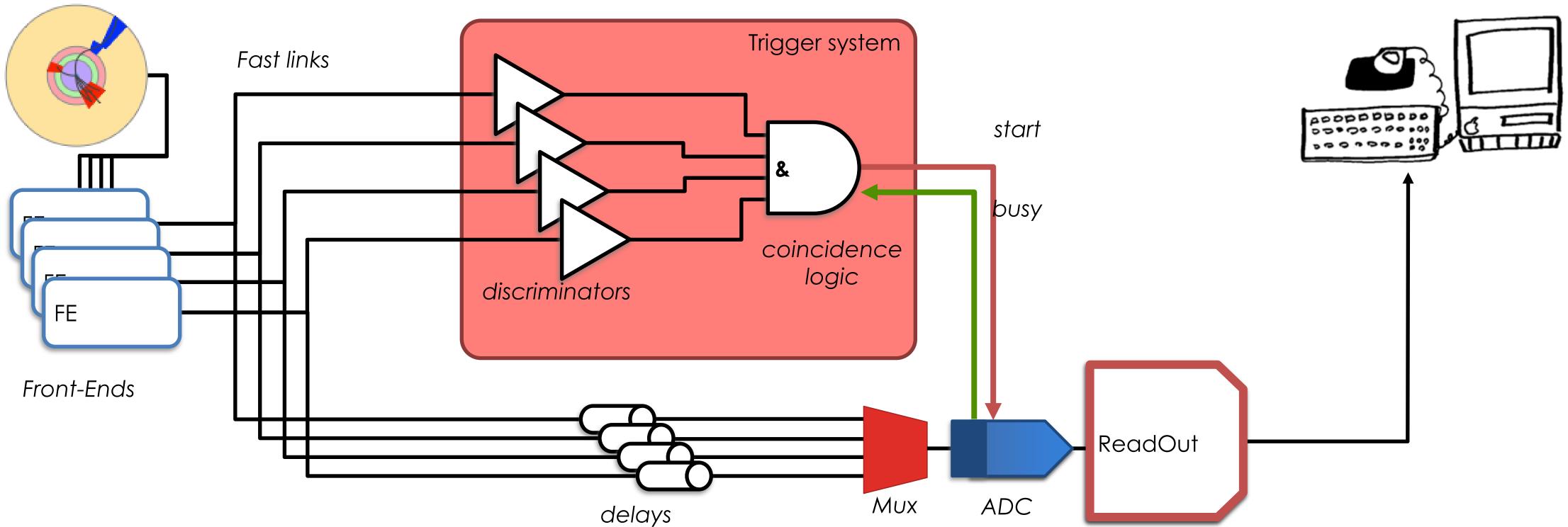
DEADTIME EXAMPLE



- Writing to disk or tape is much slow than accepting data into RAM
- If you select an event and start writing it to disk, you cannot accept any more events until you finish writing, even if they are interesting
- Same principle applies to processing time
 - For example, ADCs

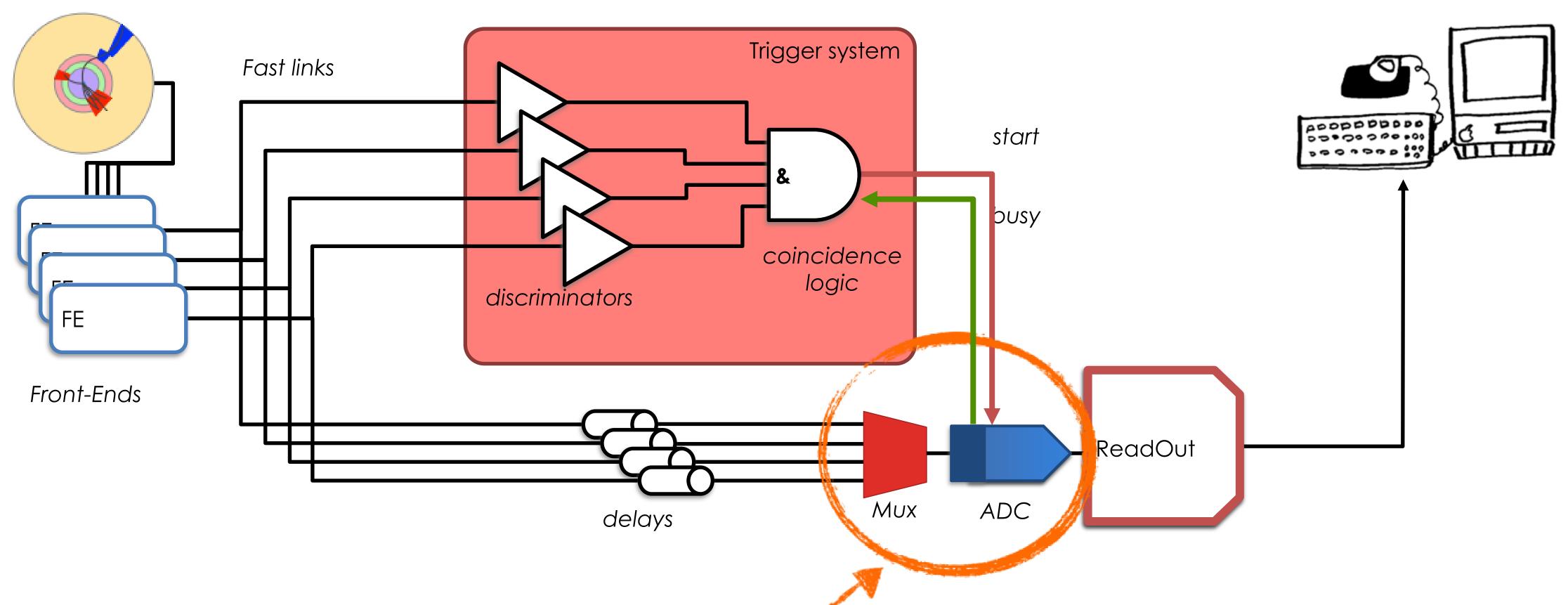
DEADTIME

A SIMPLE TRIGGER SYSTEM: DEADTIME





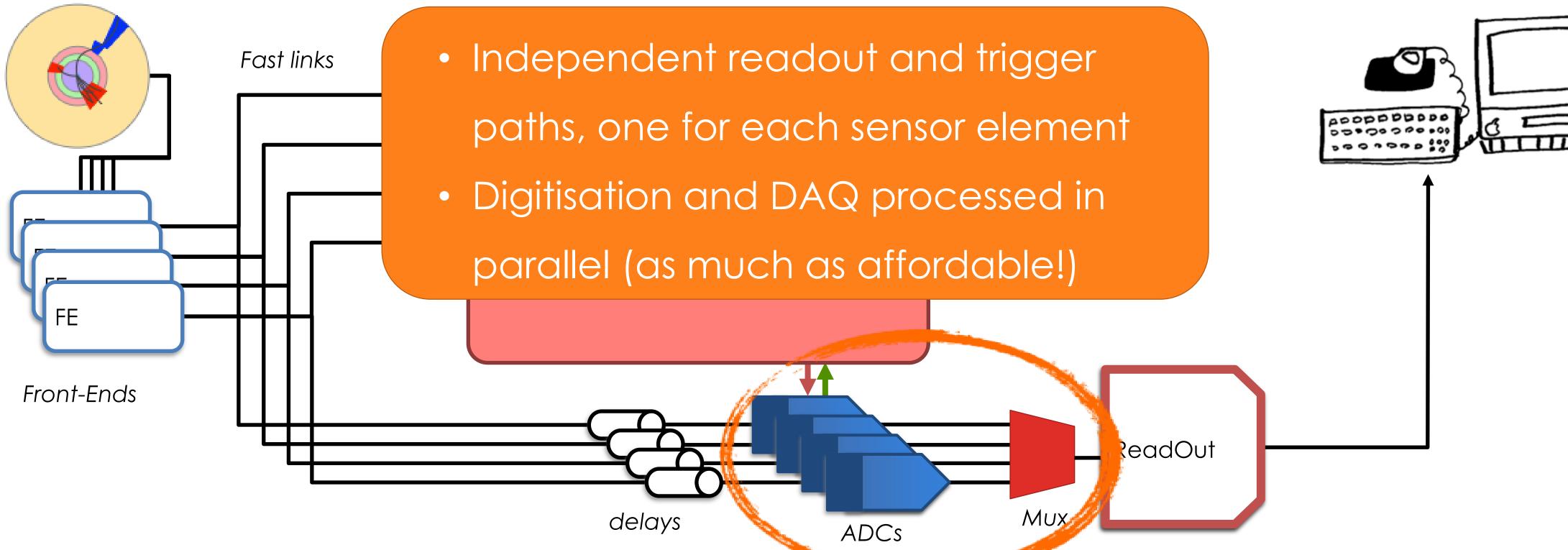
A SIMPLE TRIGGER SYSTEM: DEADTIME



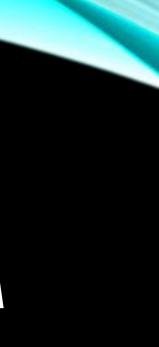
If ADC is the critical step for deadtime, this is clearly a really bad plan



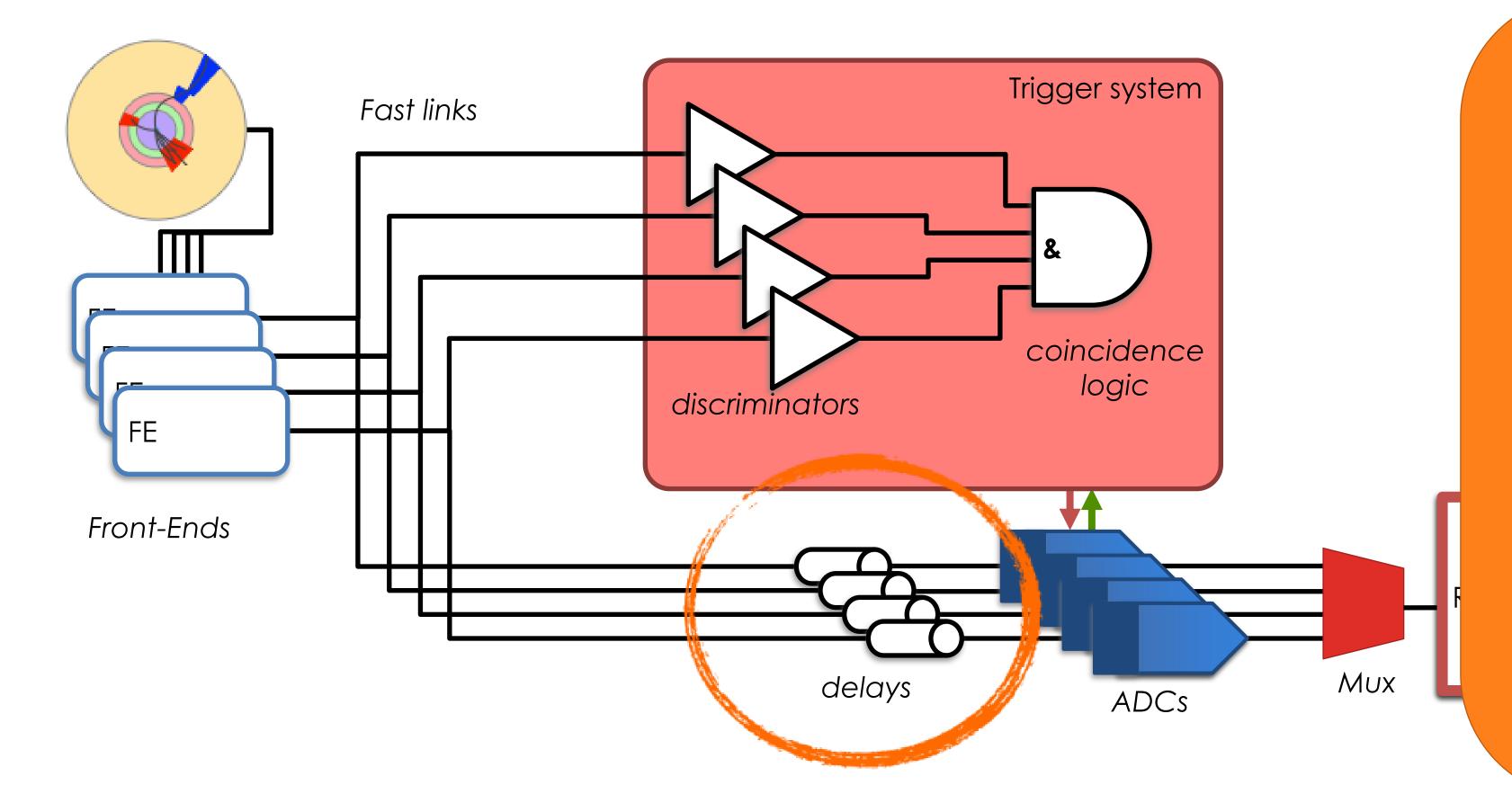
A SIMPLE TRIGGER SYSTEM: PARALLELISM



Much more sensible! Potentially much more expensive!

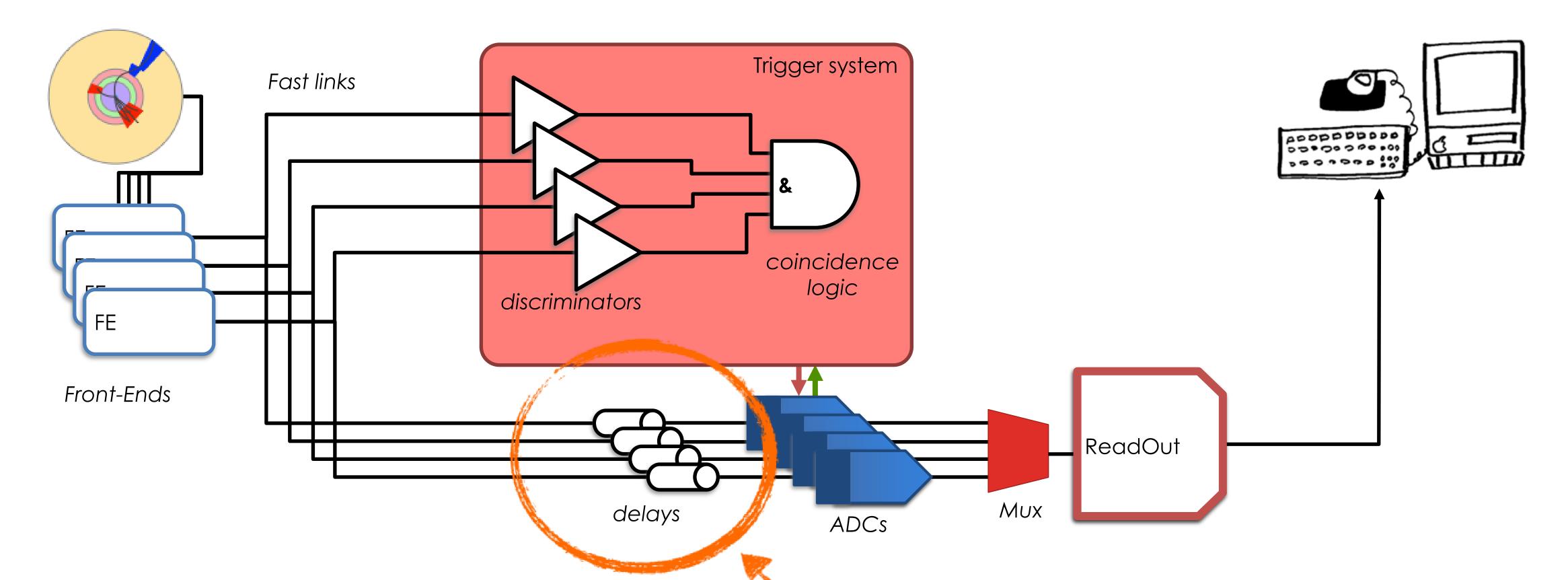


A SIMPLE TRIGGER SYSTEM: LATENCY



- Latency: Time to form the trigger decision and distribute to the digitisers
- Signals must be delayed until the trigger decision is available
- The more complex is the selection, the longer is the latency

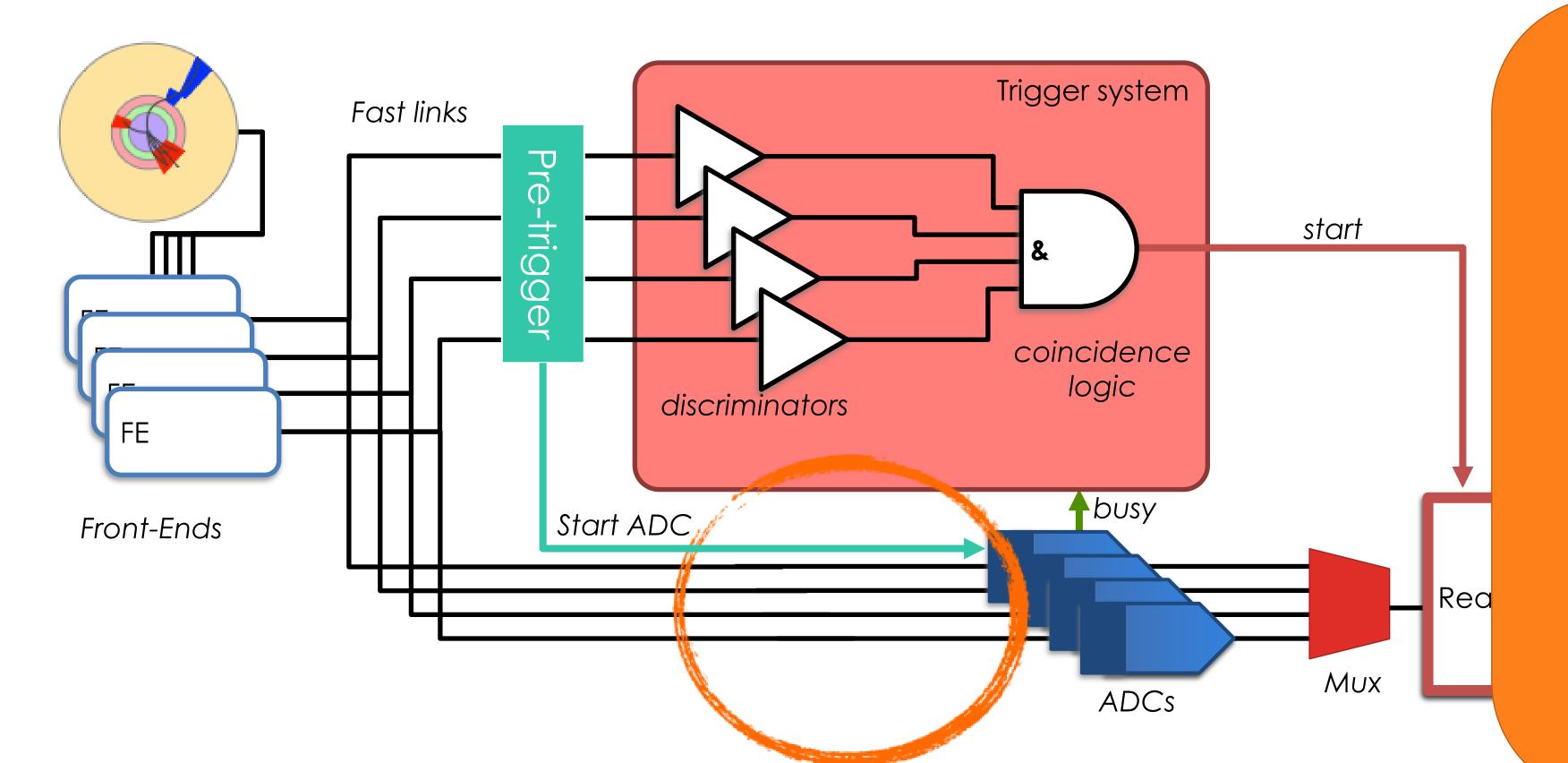
A SIMPLE TRIGGER SYSTEM: LATENCY



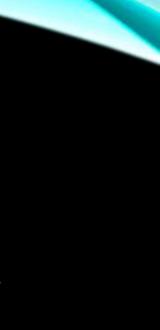
Analogue delay-lines are a bit risky, don't you think? Especially for more than one channel



A SIMPLE TRIGGER SYSTEM: PRE-TRIGGER



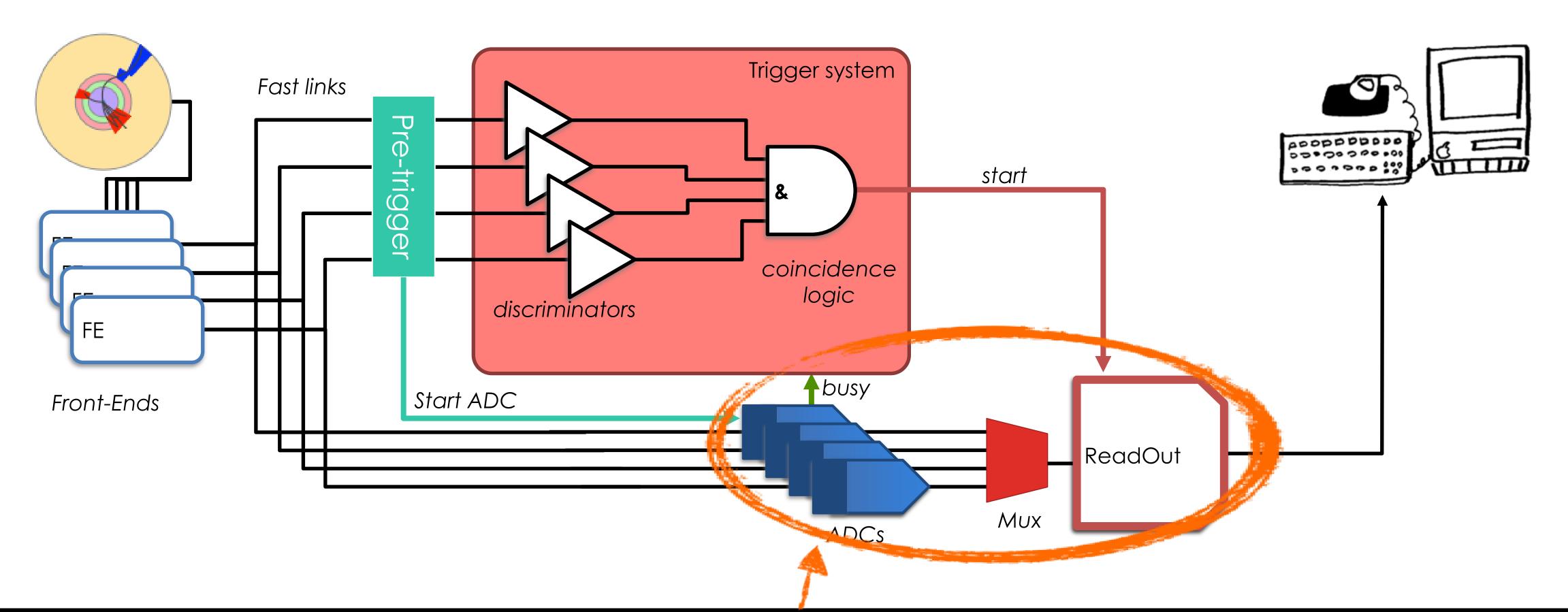
- Pre-Trigger stage: very fast indicator of some minimal activity in the detector
- Used to START the digitisers, with no delay
- The complex trigger decision comes later







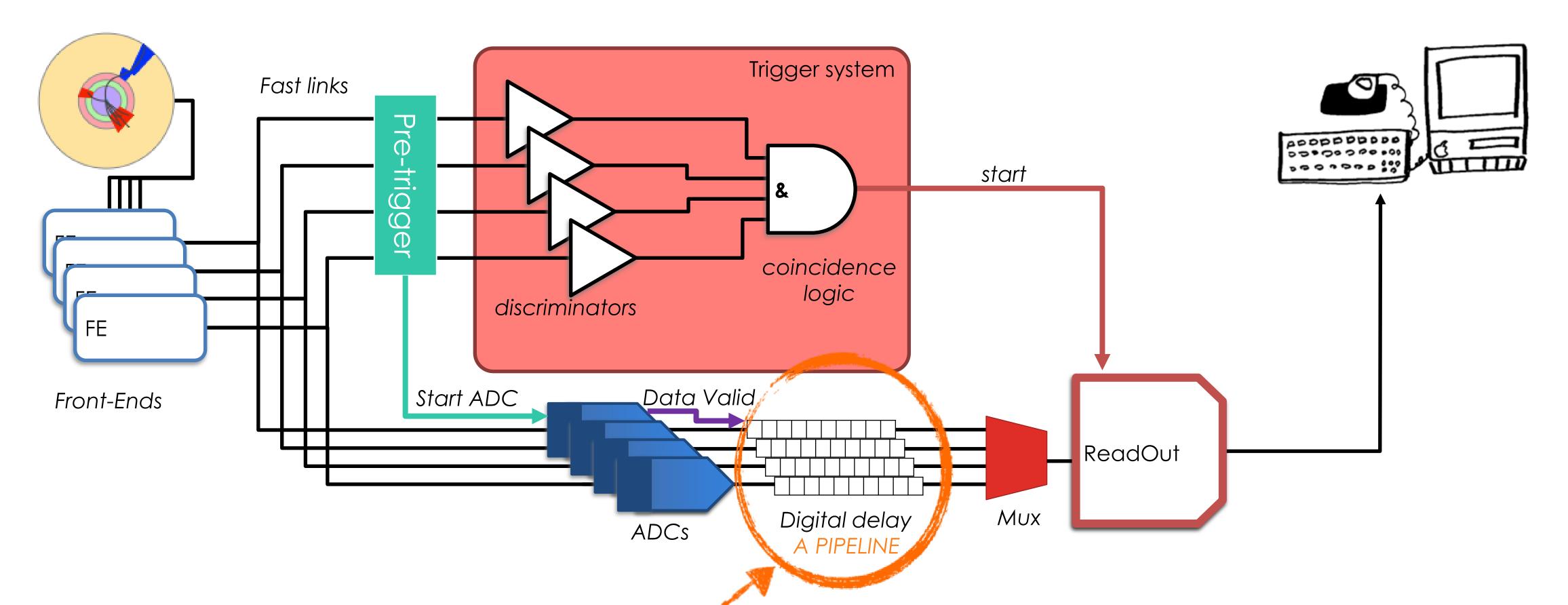
A SIMPLE TRIGGER SYSTEM: PRE-TRIGGER



Assumes the digitization time is longer than the latency of the trigger system! What if that is not true?



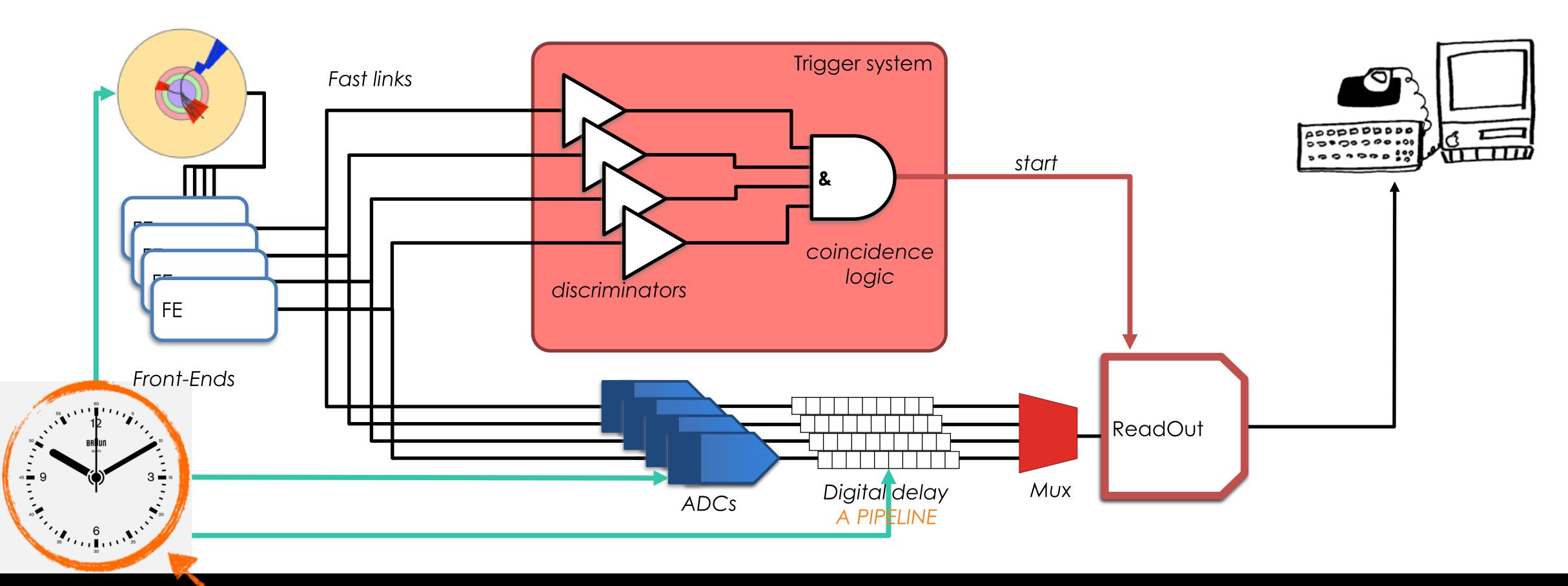
A SIMPLE TRIGGER SYSTEM: PRE-TRIGGER



Since each digitization takes a finite time Can store the result of each digitization in RAM until trigger decision is made



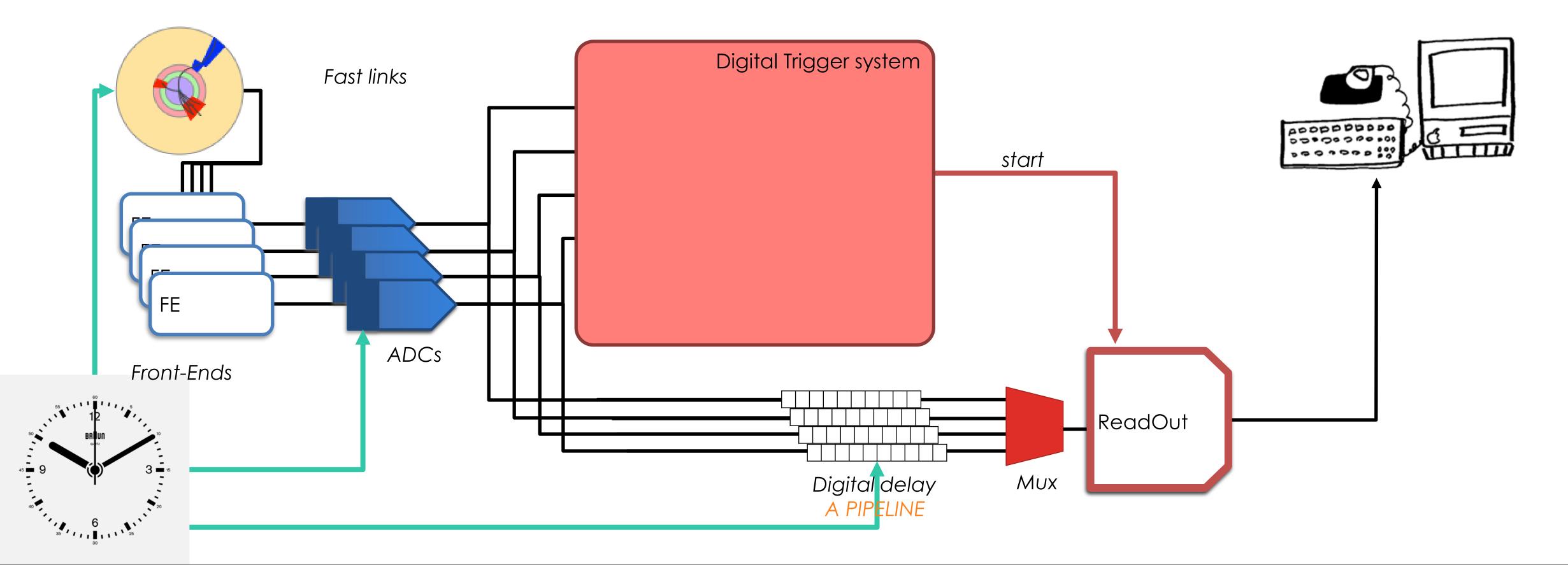
SIMPLE TRIGGER SYSTEM: BUNCHED COLLIDERS



We have a master-clock – the bunch-crossings themselves! No need for a pre-trigger

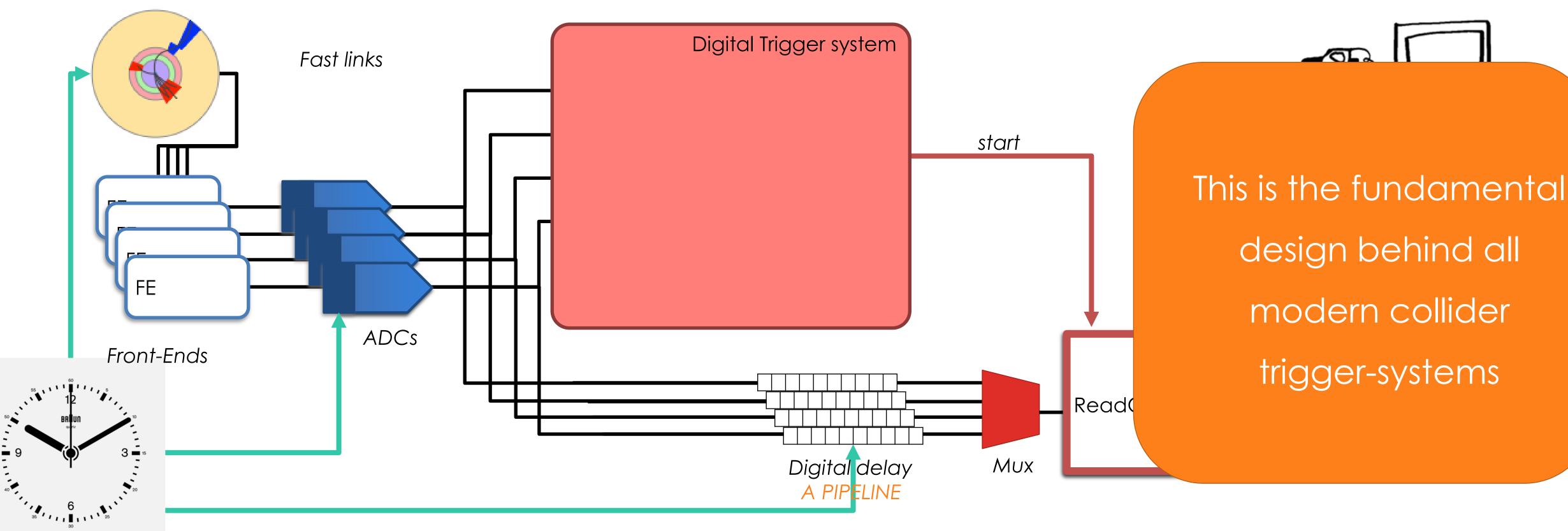


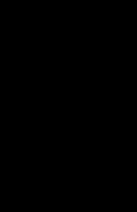
A SIMPLE TRIGGER SYSTEM: DIGITAL TRIGGERS





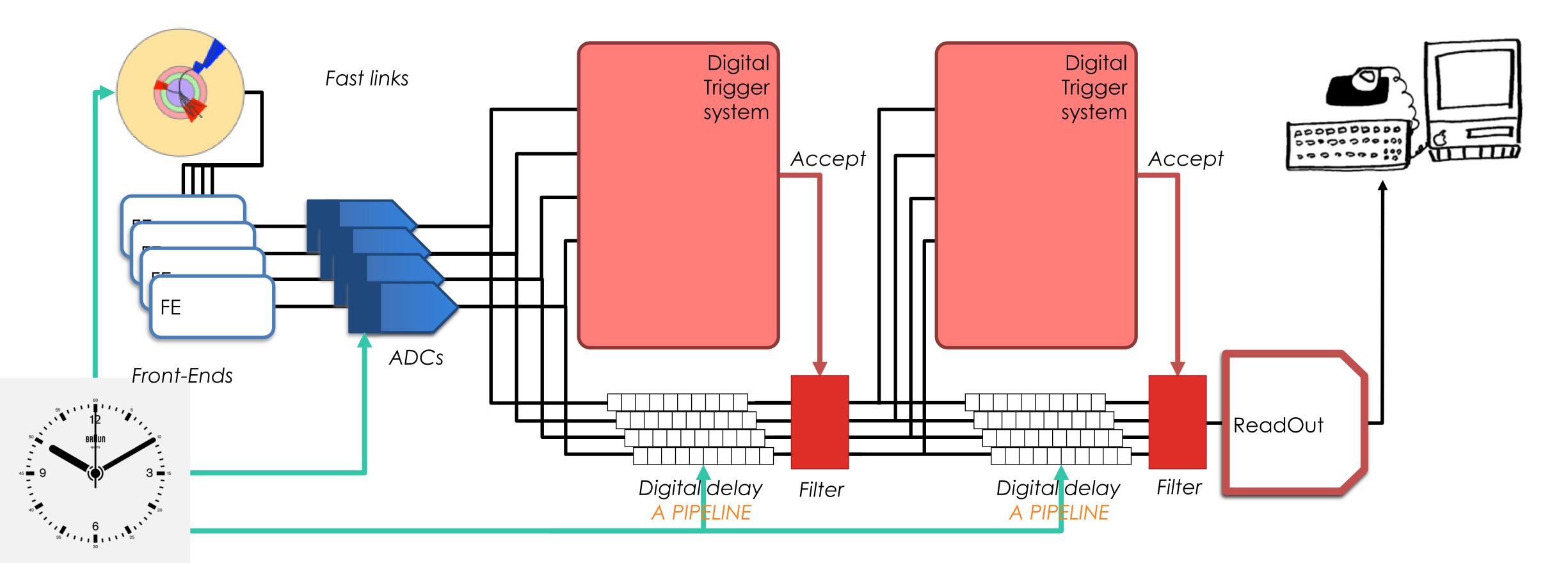
A SIMPLE TRIGGER SYSTEM: DIGITAL TRIGGERS







A TRIGGER SYSTEM: MULTILAYER TRIGGERS





- Each stage reduces the rate, so later stages have longer latency
- Complexity of algorithms increases at each level
- Dead-time is the sum of the trigger dead-time, summed over the trigger levels, and the readout dead-time

MULTILAYER TRIGGERS

- Adopted in large experiments
 - More and more complex algorithms are applied on lower and lower data rates
- Efficiency for the desired physics must be kept high AT ALL LEVELS, since rejected events are lost for ever

MULTILAYER TRIGGERS

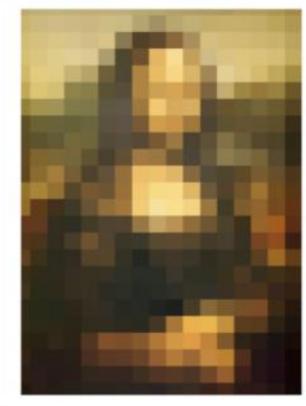
Level-1

Level-2









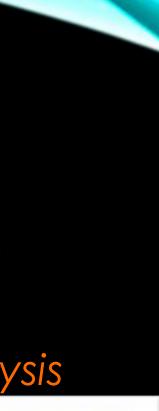




- Low latency
- Full event rate
- Small event fragment size
- Lower algorithmic complexity
- Access to coarse granularity information

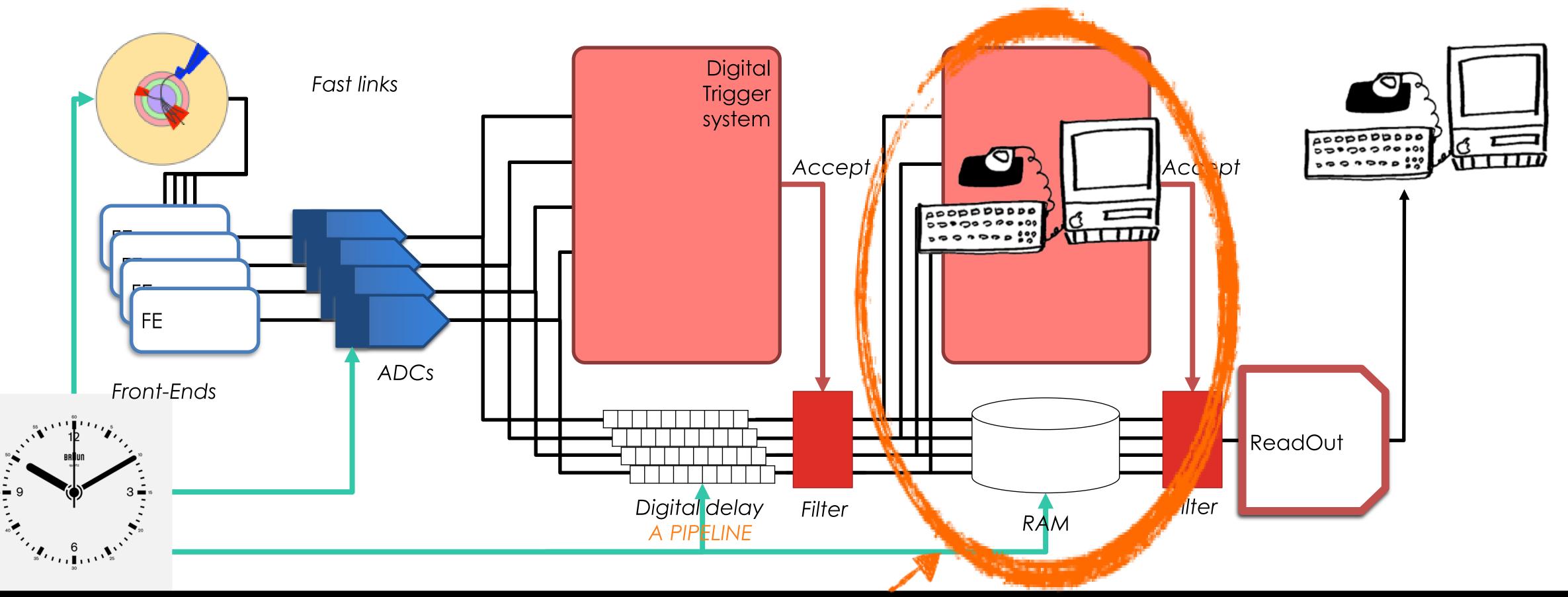
LHC experiments @ Run1		
Experiment	Number of Levels (excl. analysis)	
ATLAS	3	
CMS	2	
LHCB	3	
ALICE	4	

- Longer latency
- Lower event rate
- Larger event fragment size
- Higher algorithmic complexity
- Access to higher granularity information





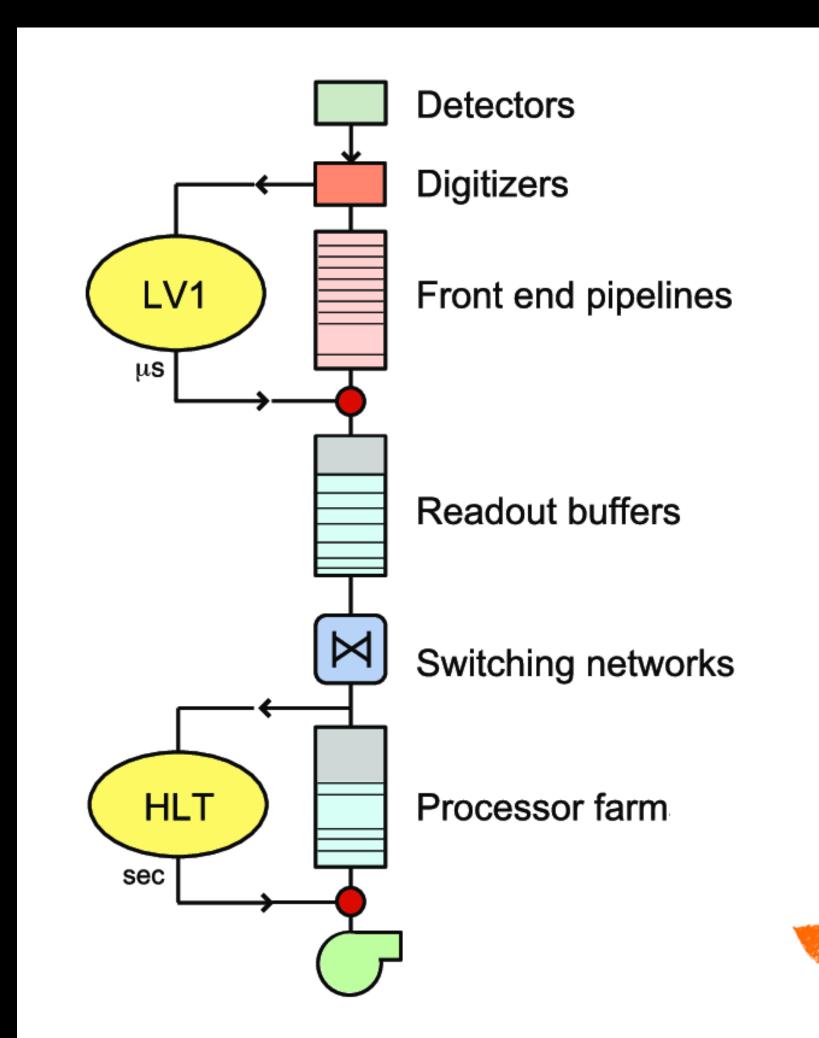
A TRIGGER SYSTEM: MULTILAYER TRIGGERS



If your input rate is low enough



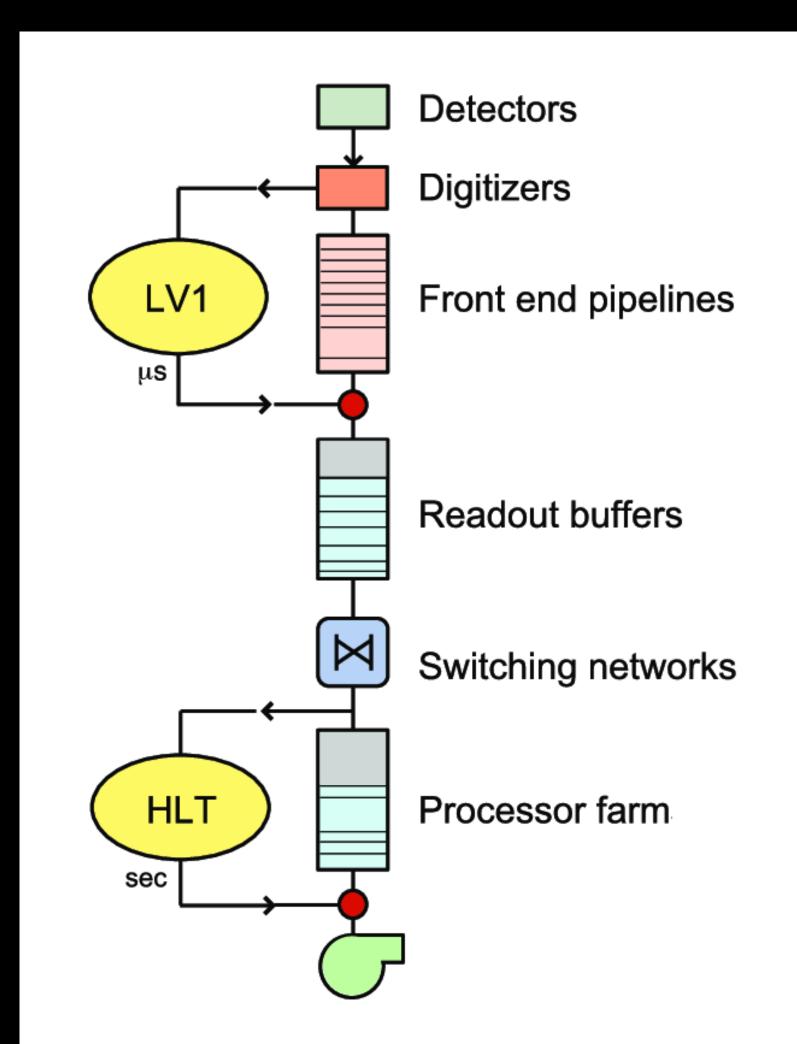
A TRIGGER SYSTEM: MULTILAYER TRIGGERS

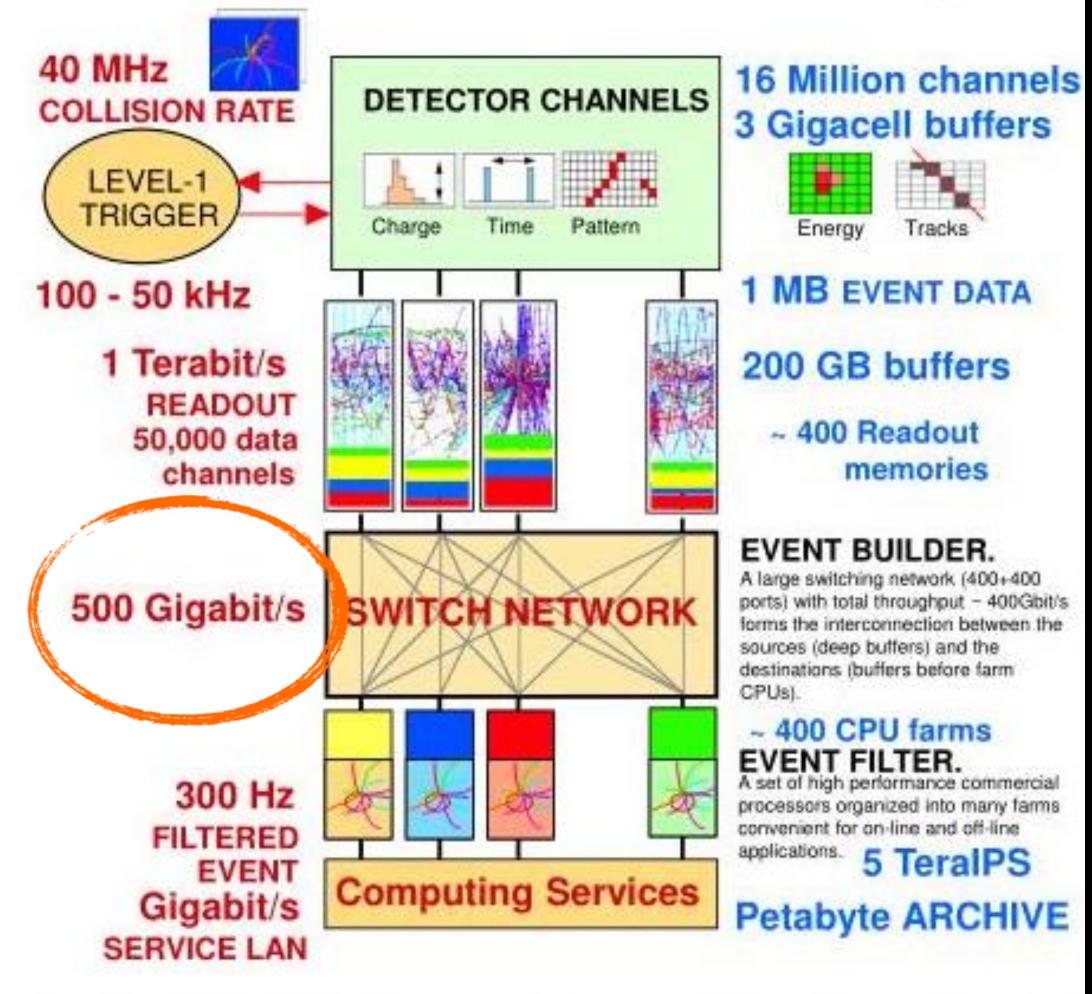


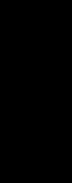
" "Standard" figure for the CMS Trigger & DAQ

 And this is exactly what the CMS Trigger does

OF COURSE, "LOW ENOUGH" IS RELATIVE...





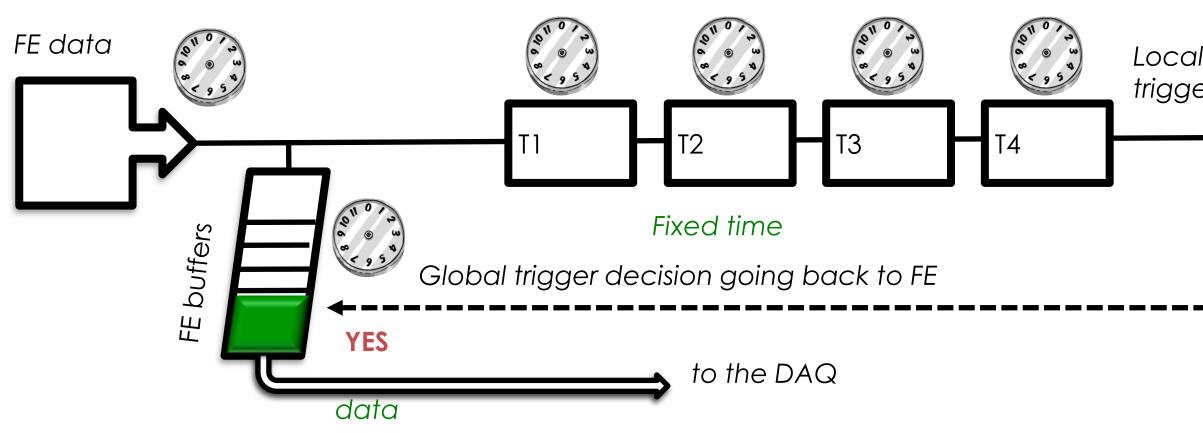


SYNCHRONOUS OR ASYNCHRONOUS?

- Synchronous: operates phase-locked with master clock
 - Data move in lockstep with the clock through the trigger chain
 - Fixed latency
 - The data, held in storage pipelines, are either sent forward or discarded
 - Used for L1 triggers in collider experiments, exploiting the accelerator bunch crossing clock

Pro's: dead-time free (just few clock cycles to protect buffers)

X con's: cost (high frequency stable electronics, sometimes needs to be custom made); maintain synchronicity throughout the entire system, complicated alignment procedures if the system is large (software, hardware, human...)



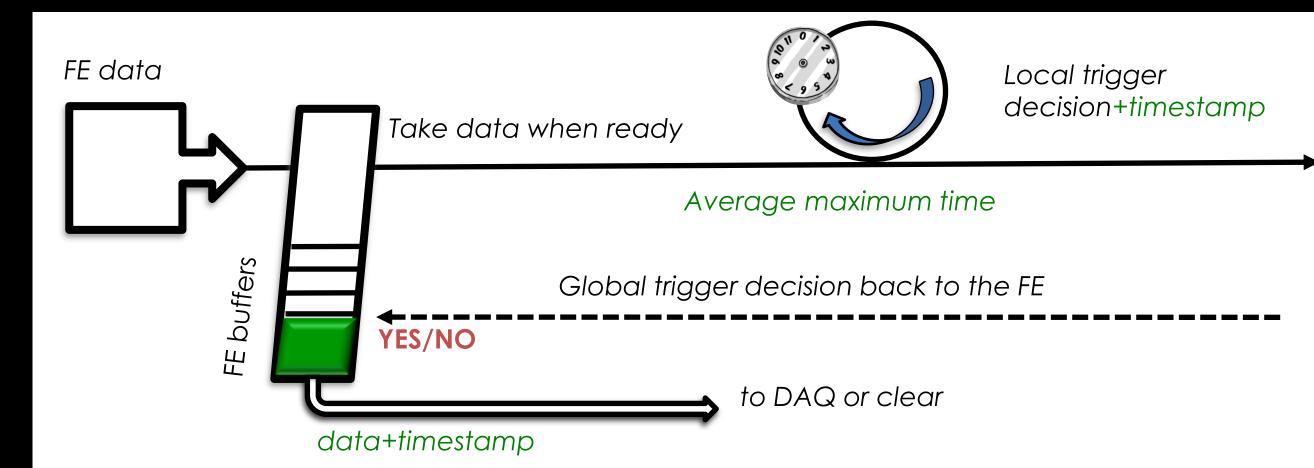
trigger decision

SYNCHRONOUS OR ASYNCHRONOUS?

- Asynchronous: operations start at given conditions (when data ready or last processing is finished)
 - Used for larger time windows \bullet
 - Average latency (with large buffers to absorb fluctuations)
 - If buffer size \neq dead-time \rightarrow lost events
 - Used also for "software filters"

Pro's: more resilient to data burst; running on conventional CPUs

Con's: needs a timing signal synchronised to the FE to latch the data, needs time-marker stored in the data, data transfer protocol is more complex)

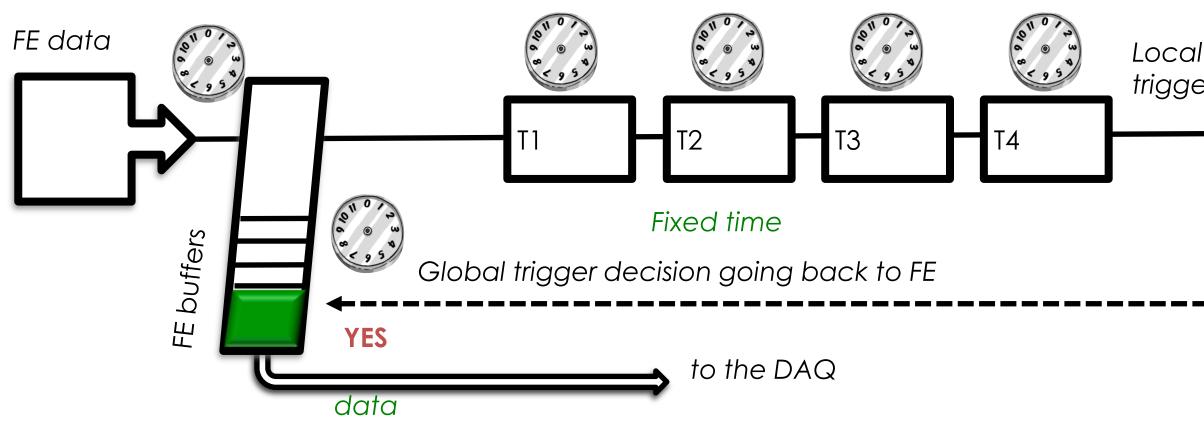


SYNCHRONOUS OR ASYNCHRONOUS? WHY NOT BOTH?

- Pseudo-synchronous: operates locally phase-locked
 - Data move in lockstep through the trigger chain from a set of local clocks
 - Buffering required whenever you move between clocks
 - Clocks run slightly faster than source data to prevent overflow
 - Realignment to global clock only after the final trigger stage
 - Fixed latency

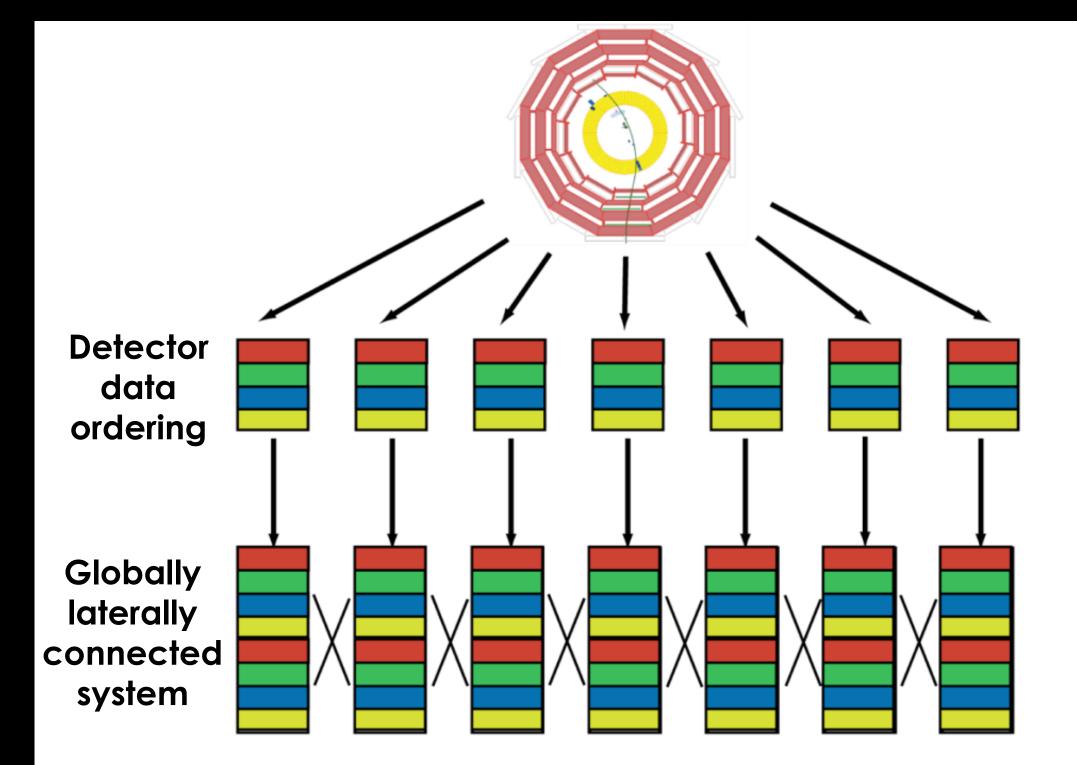
Pro's: dead-time free (just few clock cycles to protect buffers), no need for expensive globally-distributed clock, simpler alignment procedure

X con's: must propagate timing info with data, buffering required to handle clock-domain change



trigger decision

CONVENTIONAL ARCHITECTURE

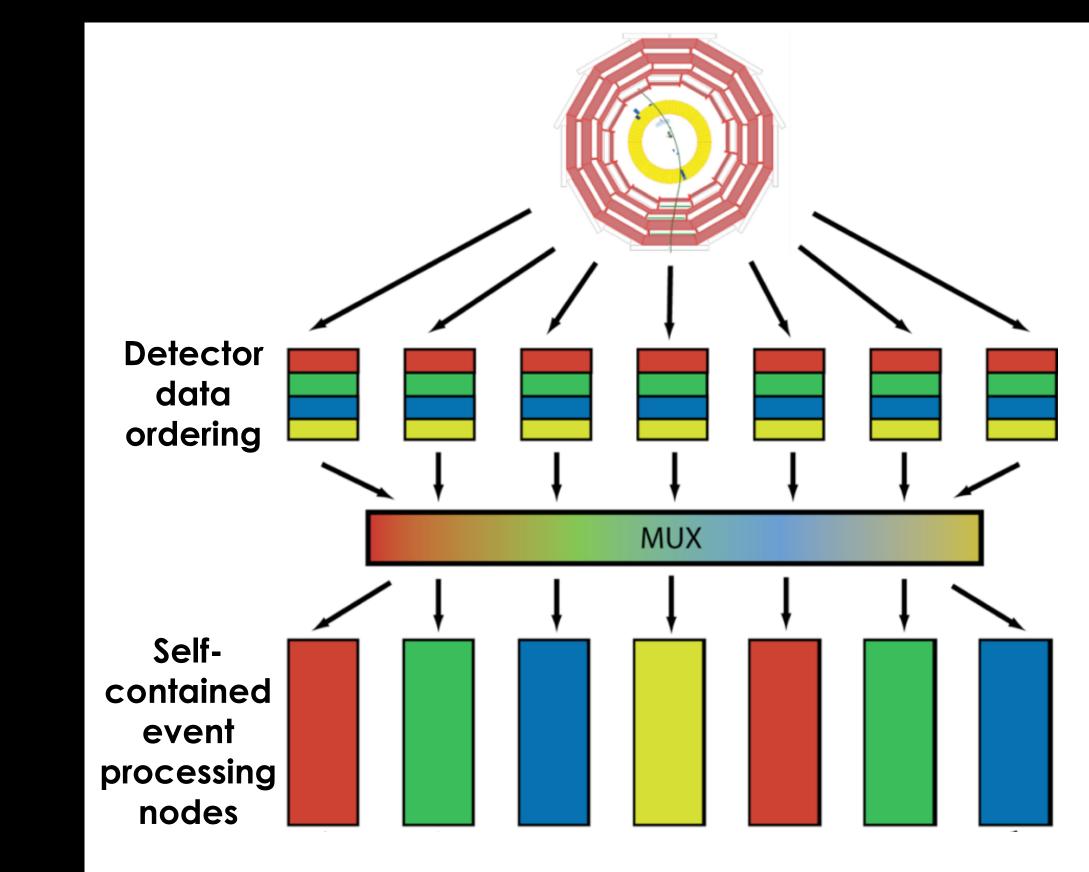


Many, many details on time-multiplexing and conventional architectures in sections 1-3 of https://cds.cern.ch/record/1421552/files/IN2011_022.pdf (although please note that the systems proposed in section 4-9 are very outdated and should be ignored)

- Each subsystem is regionally segmented
- Each region must talk to its neighbour
 - This is the root cause of requiring specialized boards for a given task!
- Each region of each processing layer compresses, suppresses, summarizes or otherwise reduces its data and passes it on to the next level which is less regionally segmented

TIME-MULTIPLEXED ARCHITECTURE

- Buffer data and stream it out optimized for processing
- Spread processing over time
 - Stream-processing rather than combinatorial-logic
 - Maximise reuse of logic resources
 - Easiest for FPGA design tools to route and meet timing
- Costs you latency, bought back by ightarrowmore efficient processing



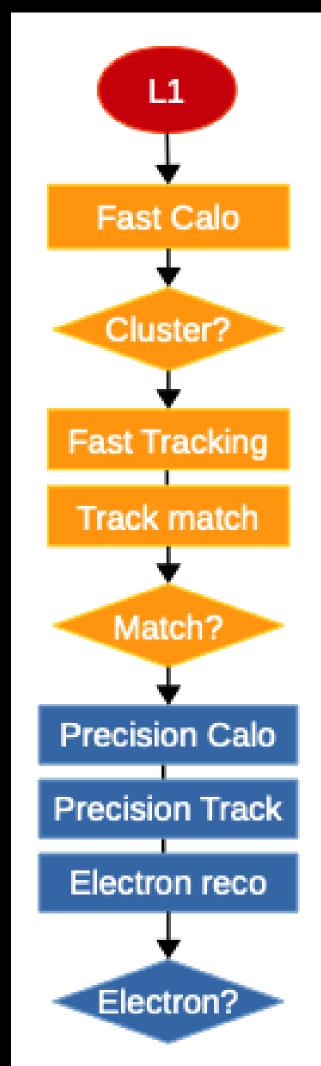
Many, many details on time-multiplexing and conventional architectures in sections 1-3 of https://cds.cern.ch/record/1421552/files/IN2011_022.pdf (although please note that the systems proposed in section 4-9 are very outdated and should be ignored)





HIGH LEVEL TRIGGER DESIGN PRINCIPLES

- Offline reconstruction too slow to be used directly ightarrow
 - Takes >10s per event
 - HLT usually needs << 1s
- Instead, step-wise processing with early rejection
 - Stop processing as soon as one step fails
 - Event accepted if any of the trigger passes ightarrow
 - Add a time-out to kill the Poisson tail!
- Fast reconstruction & L1-guided regional reconstruction first ullet
- Precision reconstruction as full detector data becomes available

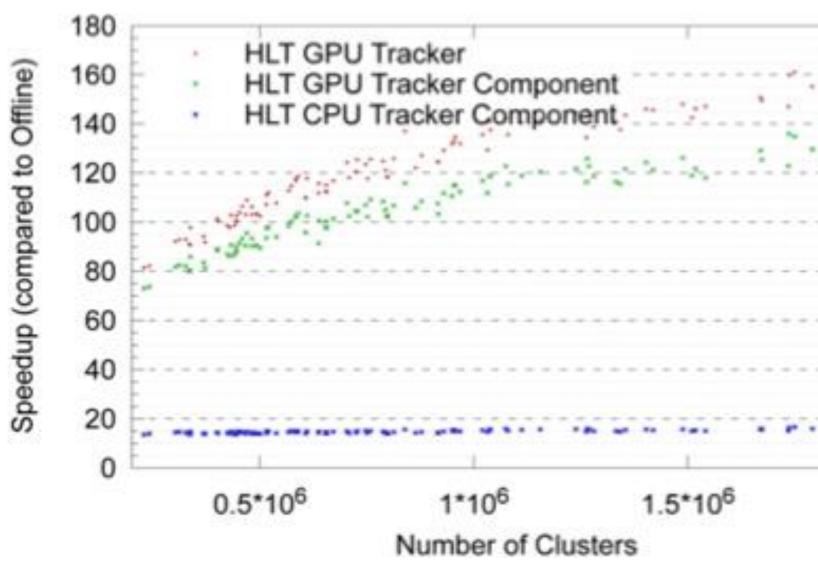


HIGH LEVEL TRIGGER DESIGN PRINCIPLES

- Early rejection reduce data and ulletresources (CPU, memory, etc.)
- Event-level parallelism
 - Process more events in parallel ullet
 - Multi-processing or/and multi-threading
- Algorithm-level parallelism ightarrow
 - **GPUs** effective whenever large amount of data ulletcan be processed concurrently (although bandwidth can be a limiting factor)

- Algorithms developed and optimized offline
- Common HLT-reconstruction software framework reduces maintenance and increases reliability







- Approximately 38,000 cores
 - An equal mix of Haswell, Broadwell and Skylake
- Multithreading allows the cores to share non-event data
 - Reduced memory footprint \rightarrow can process more events: ~20% higher performance
 - ATLAS currently doesn't have this: a race to implement this in ATHENA for Run 3
- Upgrades to add a GPU in every filter farm node is ruled out by cost and power
 - More likely a dedicated server sub-farm which does heavy tasks on demand
 - FPGAs acceleration also a (possibly better) option

EXAMPLE: CMS HLT

- Triggers are not new
 - but they are constantly evolving as the accelerators and detectors do
- FPGAs are the weapon of choice for the Level-1 trigger
 - but they are not a magic bullet
- The design of how you structure the transfer of data around your system is the most important decision you will make
- Heterogeneous computing farms look likely to feature at HL-LHC
 - but it is a brave new world!

CONCLUSION

- Triggers are not new lacksquare
 - but they are constantly evolving as the accelerators and detectors do
- FPGAs are ullet
 - but they
- important decision you will make
- Heterogeneous computing farms look likely to feature at HL-LHC \bullet
 - but it is a brave new world!

CONCLUSION

Oh, and be very suspicious if your supervisor plies you with

strong coffee and gets you to look for scintillation light

• The design of how you structure the transfer of data around your system is the most

Spares: Introduction to FPGAs



And this is not a new point

- management in order to secure efficient use.



"The parallel approach to computing does require that some original thinking be done about numerical analysis and data

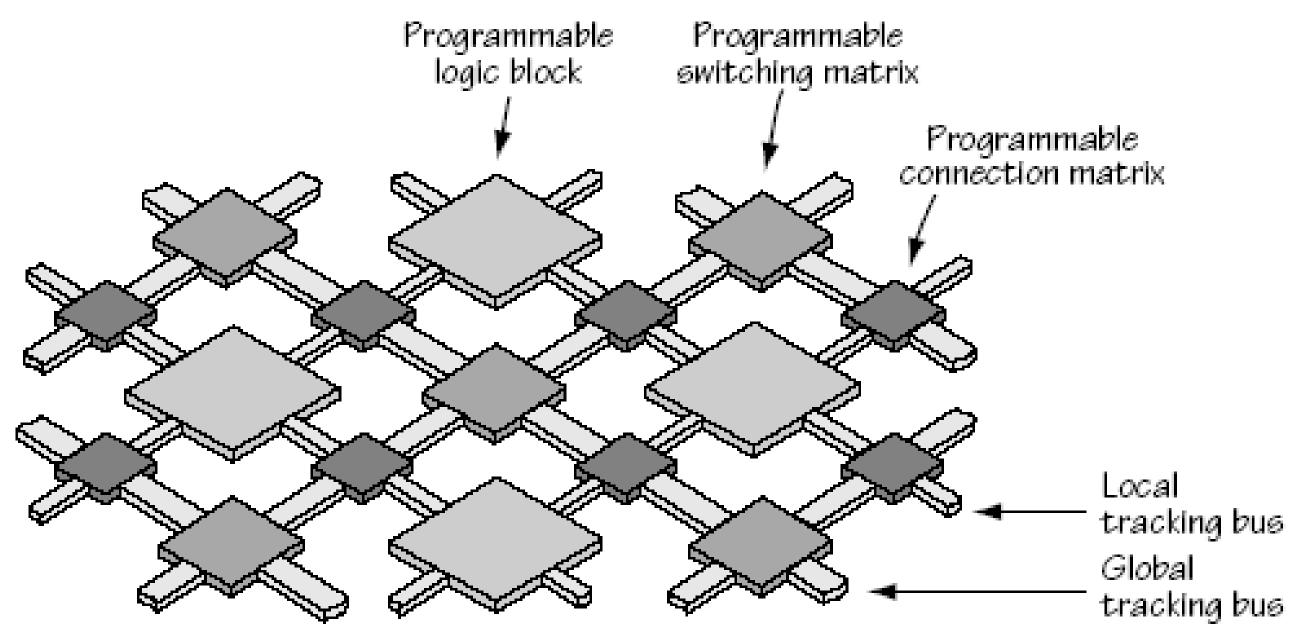
In an environment which has represented the absence of the need to think as the highest virtue, this is a decided disadvantage"

Daniel Slotnick, 1967

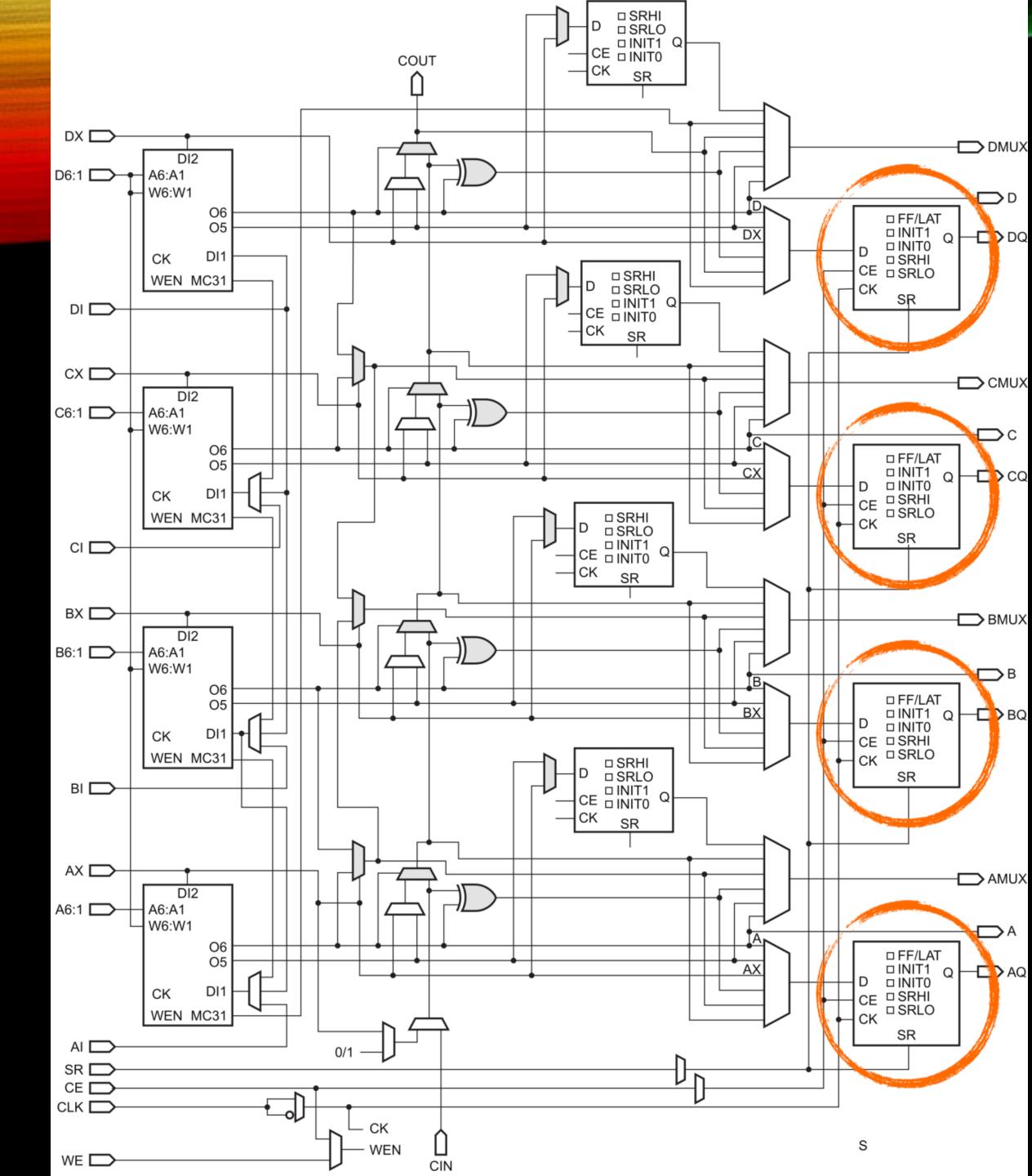


FIELD PROGRAMMABLE GATE ARRAYS

- Programmable Logic Blocks
- Massive Fabric of Programmable Interconnects



tracking bus

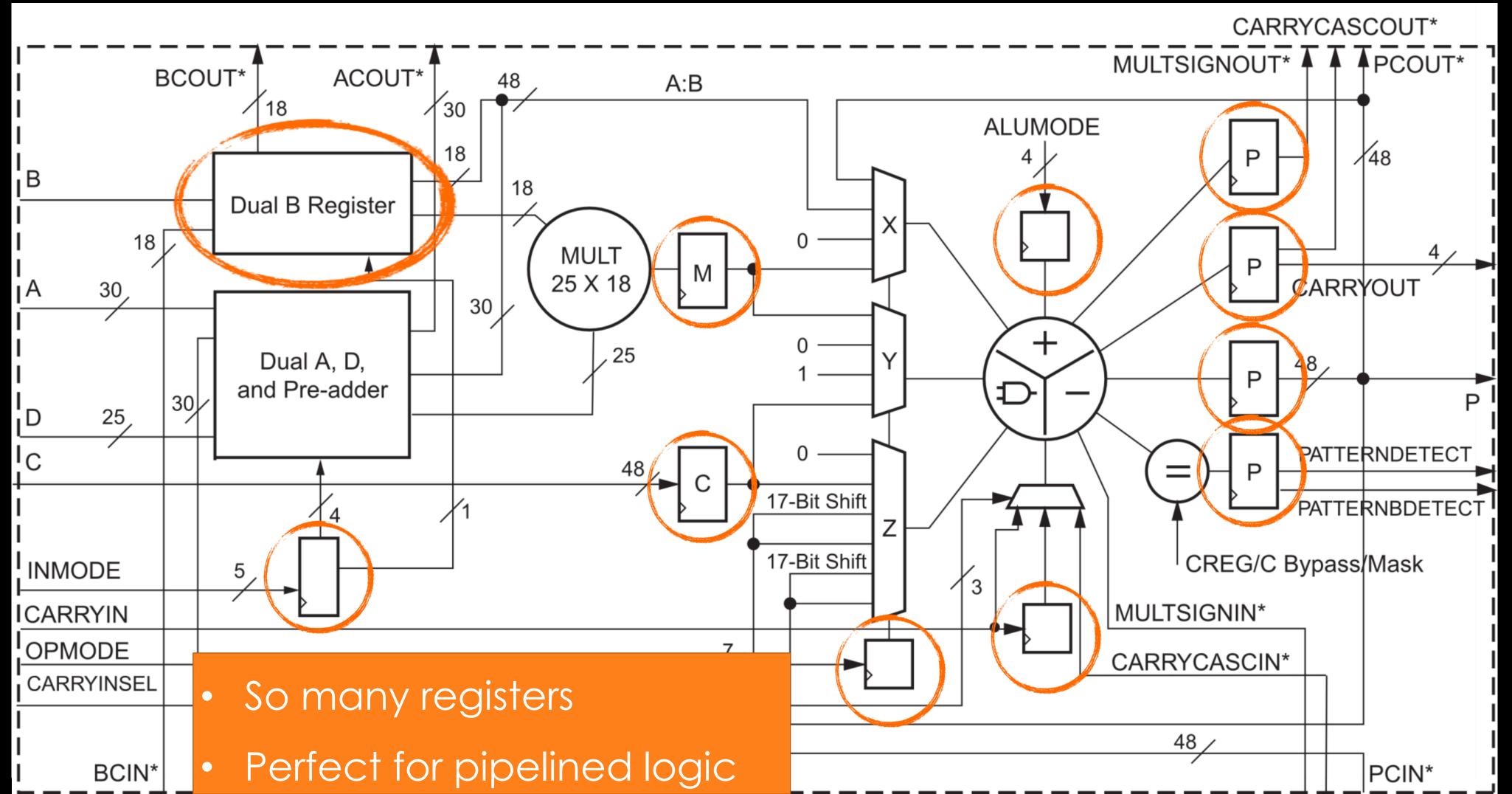


COMBINATORIAL LOGIC BLOCK

Registers on the output of every cell
Perfect for pipeline<u>d logic</u>

.

INTEGRATED DIGITAL SIGNAL PROCESSING



BIGGEST XILINX "ULTRASCALE+" DEVICES

- Upwards of 9million logic cells
 - All clocked at up to 500MHz
 - Up to $O(10^{15})$ operations per second
- Upwards of 12000 DSPs
- All pipelined
- Fully programmable
- And we have a winner!

Device Name	VU9P	VU11P	VU
Effective LEs ⁽¹⁾ (K)	2,485	2,575	3,4
Logic Cells (K)	2,069	2,147	2,8
CLB Flip-Flops (K)	2,364	2,454	3,2
CLB LUTs (K)	1,182	1,227	1,0
Max. Distributed RAM (Mb)	36.1	34.8	4(
Total Block RAM (Mb)	75.9	70.9	94
UltraRAM (Mb)	270.0	324.0	43
DSP Slices	6,840	8,928	11,
PCle® Gen3 x16 / Gen4 x8	6	3	
150G Interlaken	9	9	1
100G Ethernet w/ RS-FEC	9	6	
Max. Single-Ended HP I/Os	832	624	8
GTY 32.75Gb/s Transceivers	120	96	1



BIGGEST XILINX "ULTRASCALE+" DEVICES

- Upwards of 9 million logic cells
 - All clocked at up to 500MHz
 - Up to $O(10^{15})$ operations per second
- Upwards of 12000 DSPs
- All pipelined
- Fully programmable
- And we have a winner!
- So what is the catch?

Device Name	VU9P	VU11P	VU
Effective LEs ⁽¹⁾ (K)	2,485	2,575	3,4
Logic Cells (K)	2,069	2,147	2,8
CLB Flip-Flops (K)	2,364	2,454	3,2
CLB LUTs (K)	1,182	1,227	1,0
Max. Distributed RAM (Mb)	36.1	34.8	4(
Total Block RAM (Mb)	75.9	70.9	94
UltraRAM (Mb)	270.0	324.0	43
DSP Slices	6,840	8,928	11,
PCle® Gen3 x16 / Gen4 x8	6	3	
150G Interlaken	9	9	1
100G Ethernet w/ RS-FEC	9	6	
Max. Single-Ended HP I/Os	832	624	8
GTY 32.75Gb/s Transceivers	120	96	1



BIGGEST XILINX "ULTRASCALE+" DEVICES

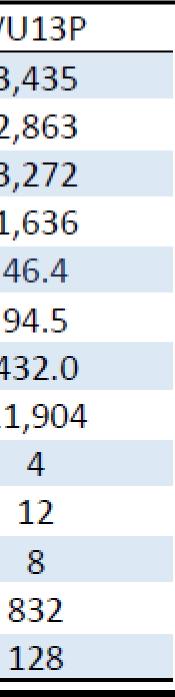
- Upwards of 9n ullet
 - All clocked
 - Up to O(10¹
- Upwards of 12
- All pipelined
- Fully programr ullet

Incredibly hard to program efficiently

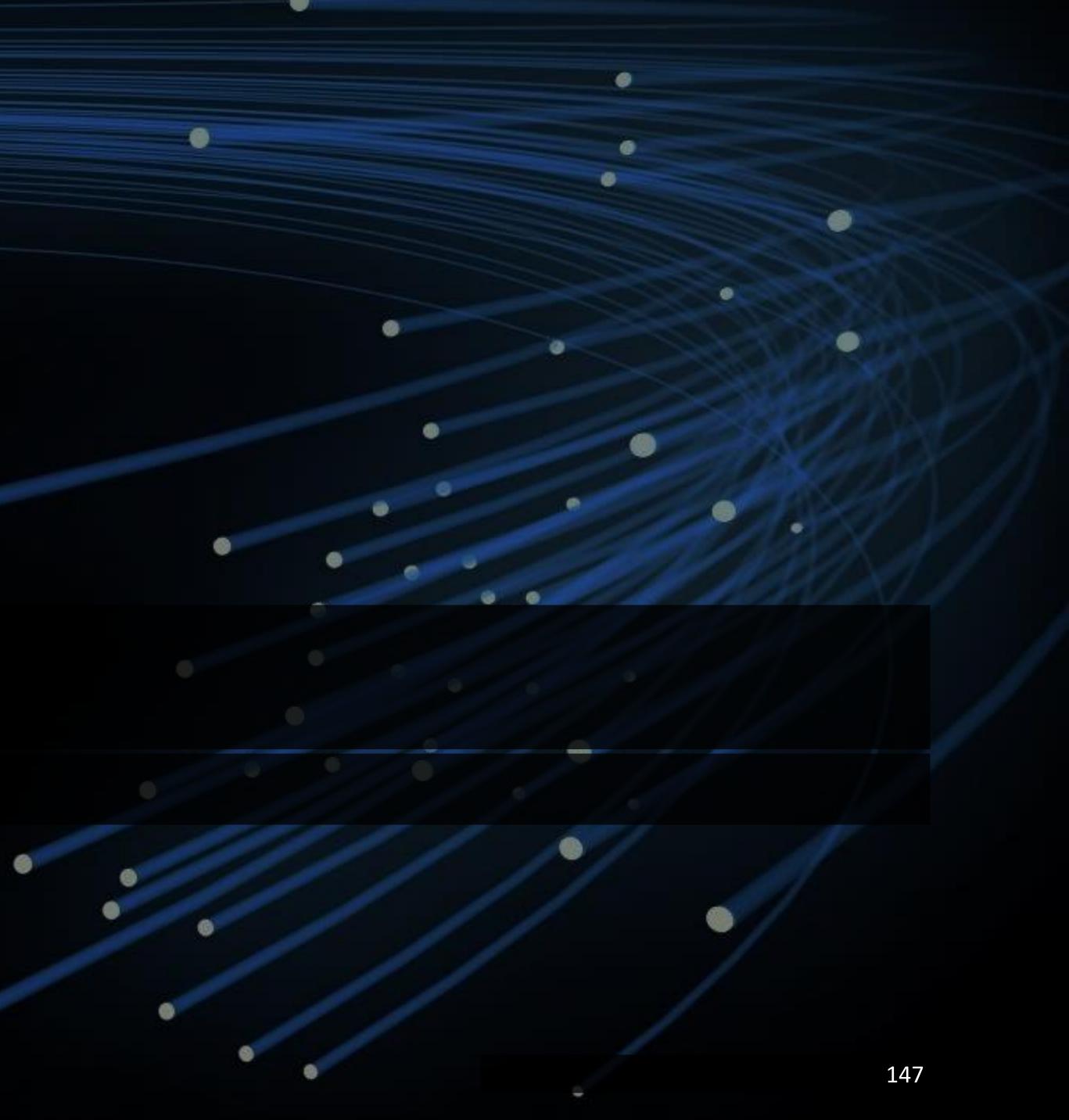
- Thinking in a parallel, pipelined-fashion is • exceptionally difficult
- A handful of real experts in CMS
- Efficient use depends on efficiently structured data
- The chip is just the start needs to be attached to something
- And we have a winner.
- So what is the catch?

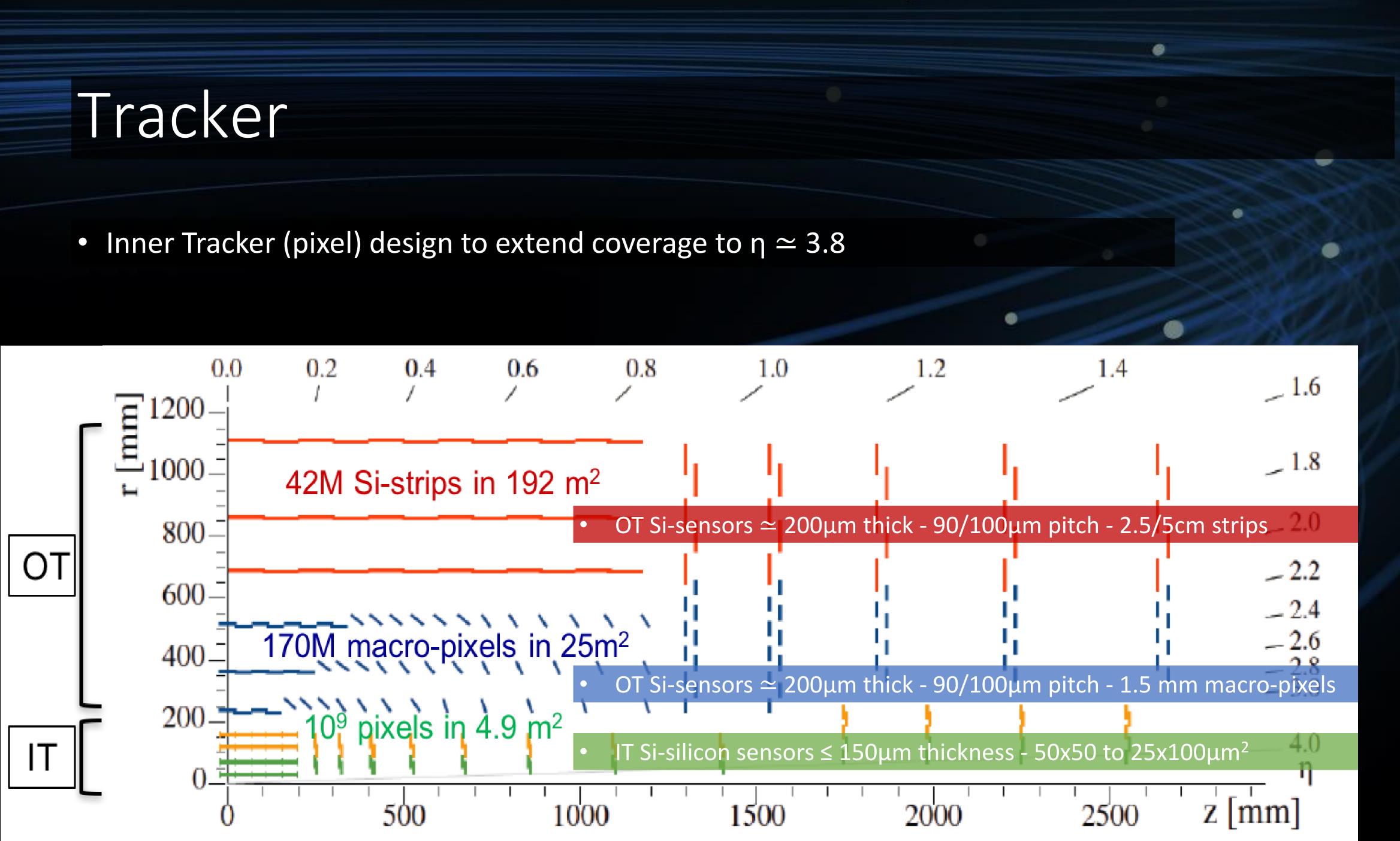
You are also responsible for the infrastructure

VU11P	V
2,575	3
2,147	2
2,454	3
1,227	1
34.8	4
70.9	9
324.0	4
8,928	11
3	
9	
6	
624	2
96	



Spares: Phase-2

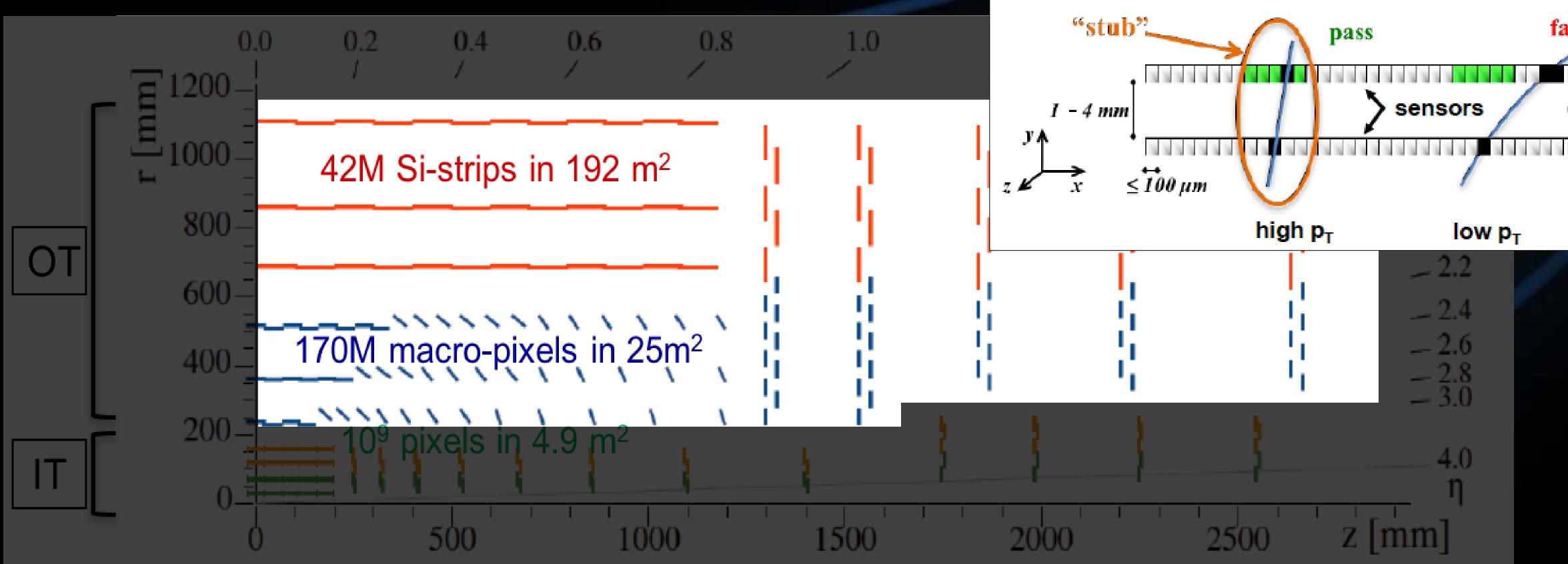




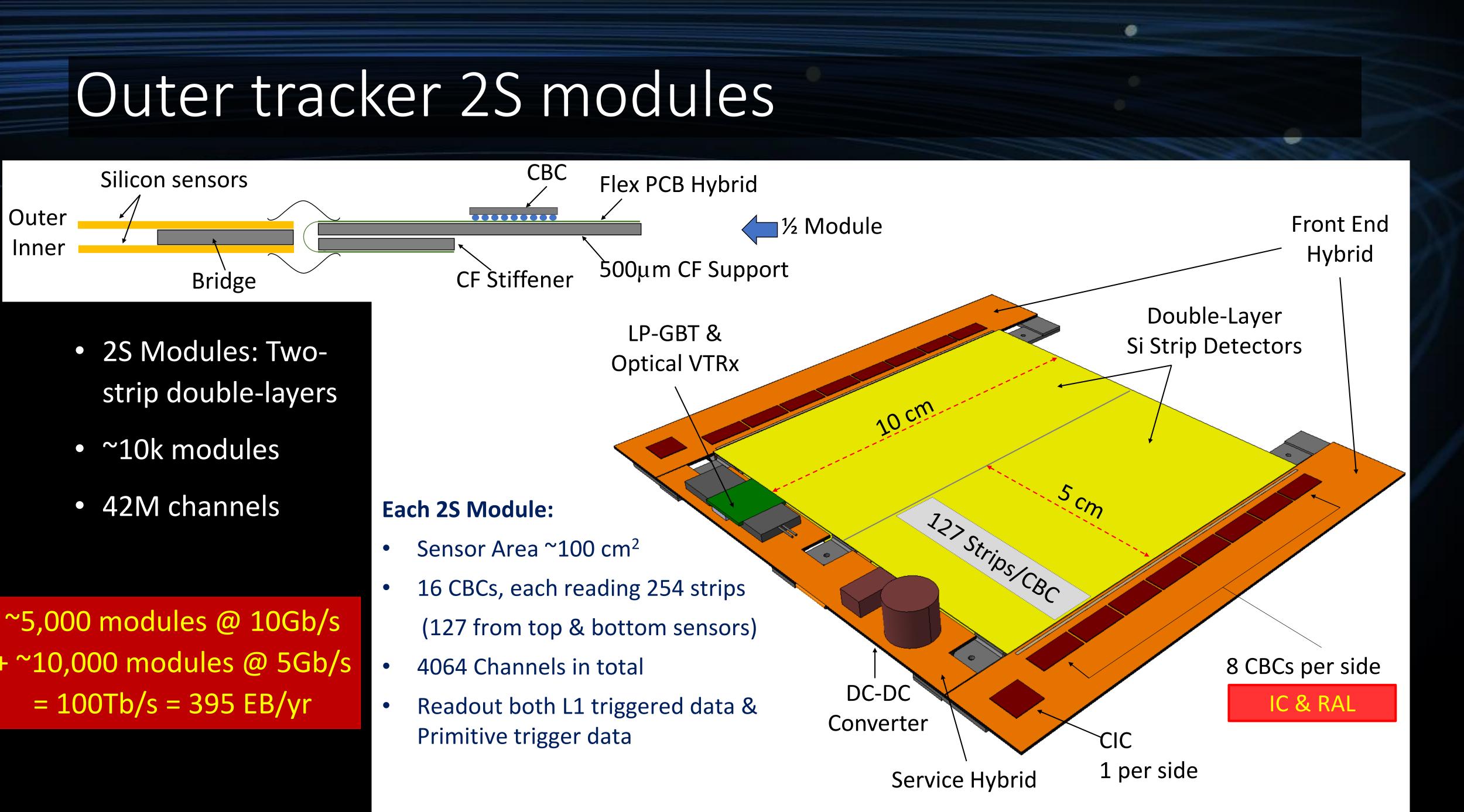


Tracker

- Inner Tracker (pixel) design to extend coverage to $\eta \simeq 3.8$
- Outer Tracker design driven by ability to provide tracks at 40 MHz to L1-trigger \bullet

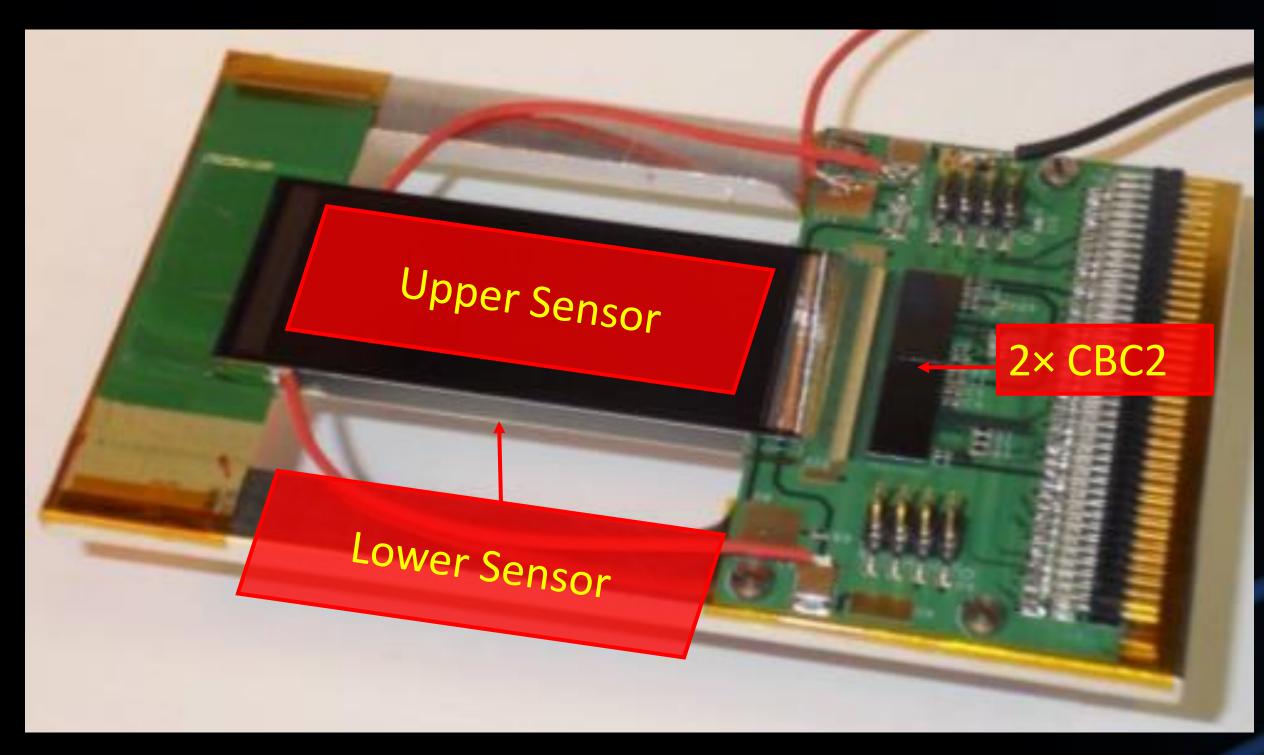


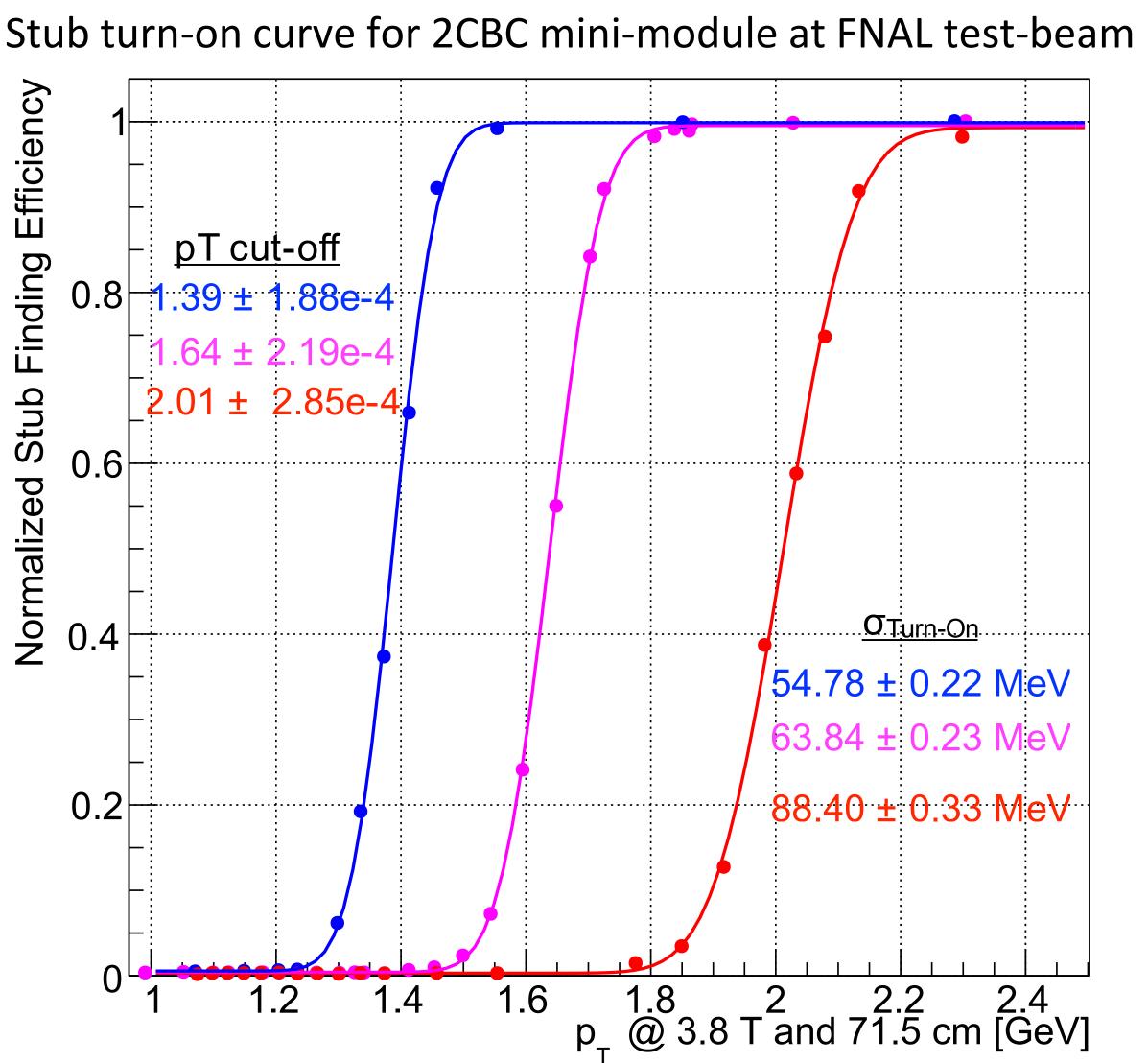




+~10,000 modules @ 5Gb/s

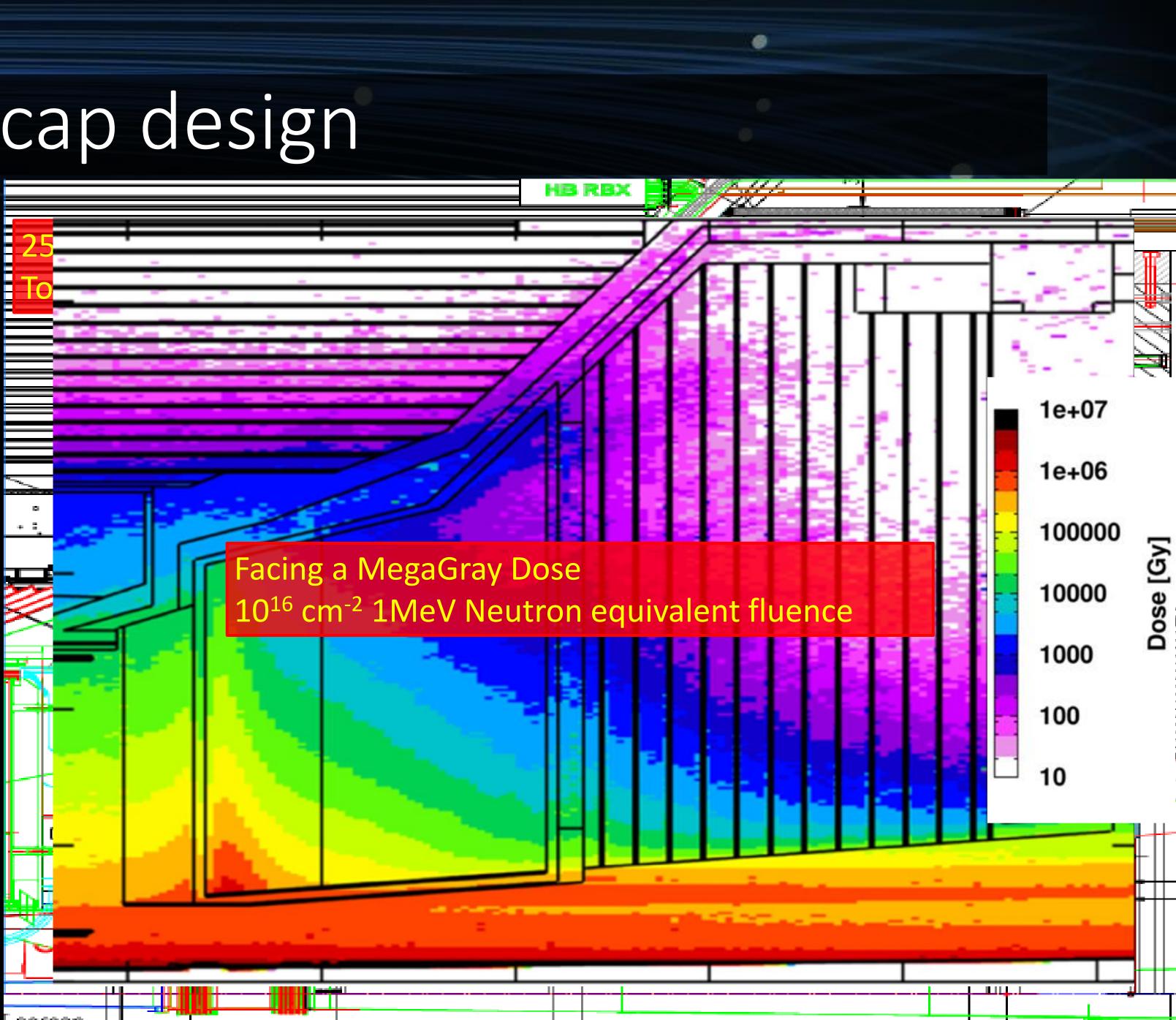
Outer tracker 2S modules: Do they work?





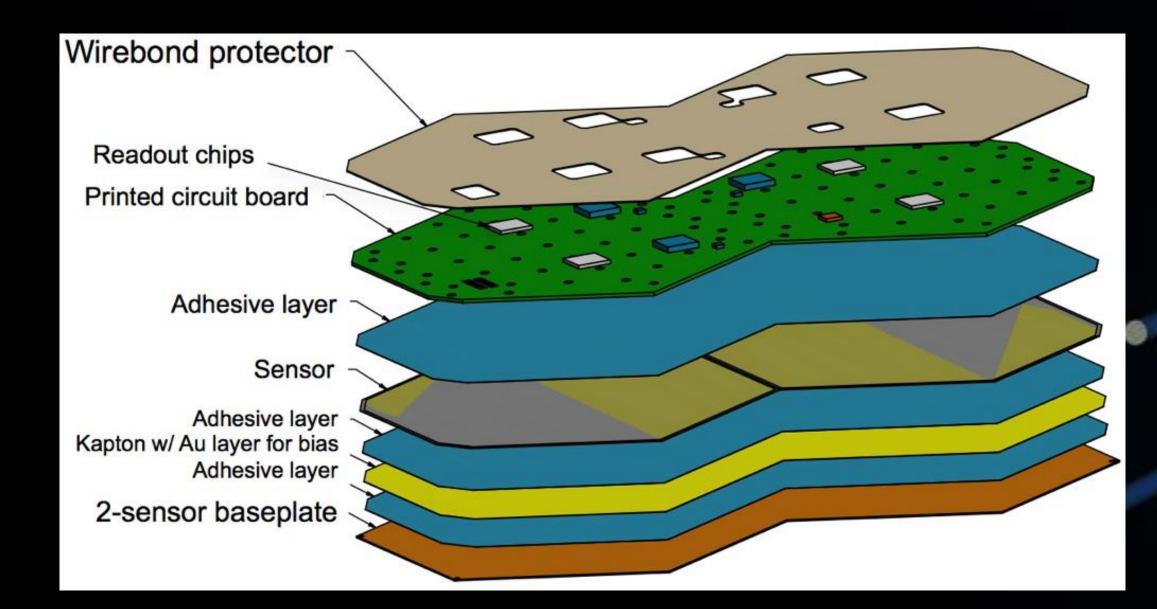
Calorimeter Endcap design

- 3D shower topology and time resolution of ~30ps
- Electromagnetic Endcap (EE)
 - 28 layers of Silicon sensors in W/Pb absorber (25 X₀, 1.7λ)
- Hadronic Endcap (EH)
 - 24 layers: 8 silicon + 16 silicon/scint. tiles at high/low η in stainless steel absorber (9λ)



Calorimeter Endcap modules

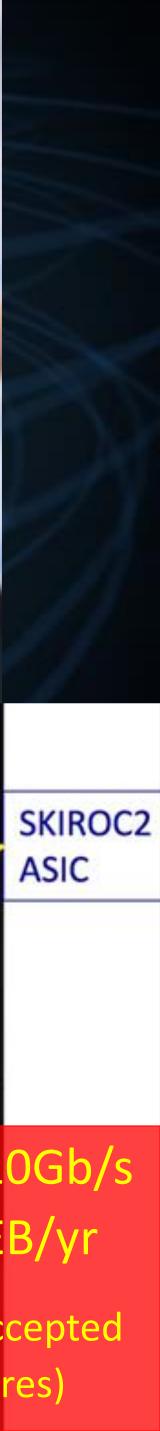
- 593 m2 of silicon
- 6M ch, 0.5 or 1 cm2 cell-size
- 21,660 modules (8" or 2x6" sensors)
- 92,000 front-end ASICS



9,670 modules @ 10Gb/s = 100Tb/s ≈ 395 EB/yr

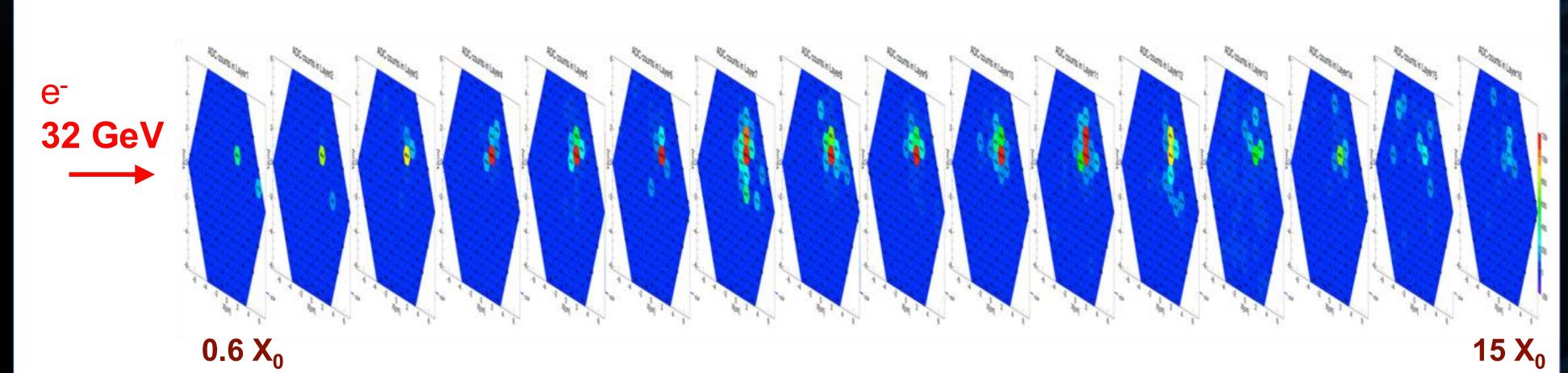
2.0

(Trigger data only, L1 accepted data on separate fibres)

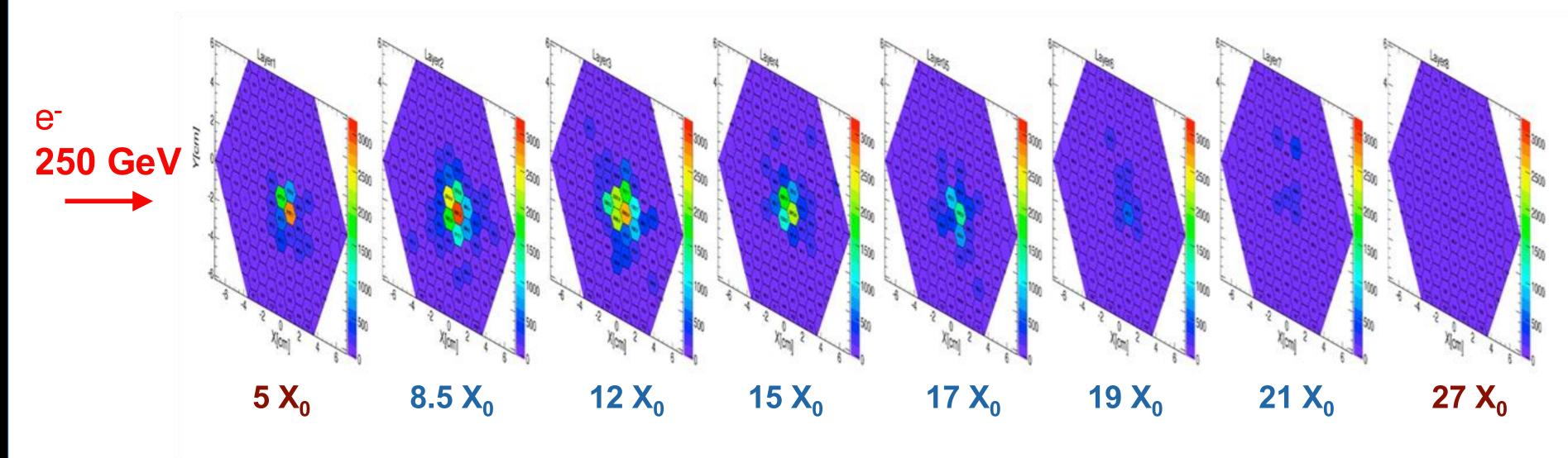


Calorimeter Endcap modules: Do they work?

Fermilab: 32 GeV electrons passing through 15 X₀.



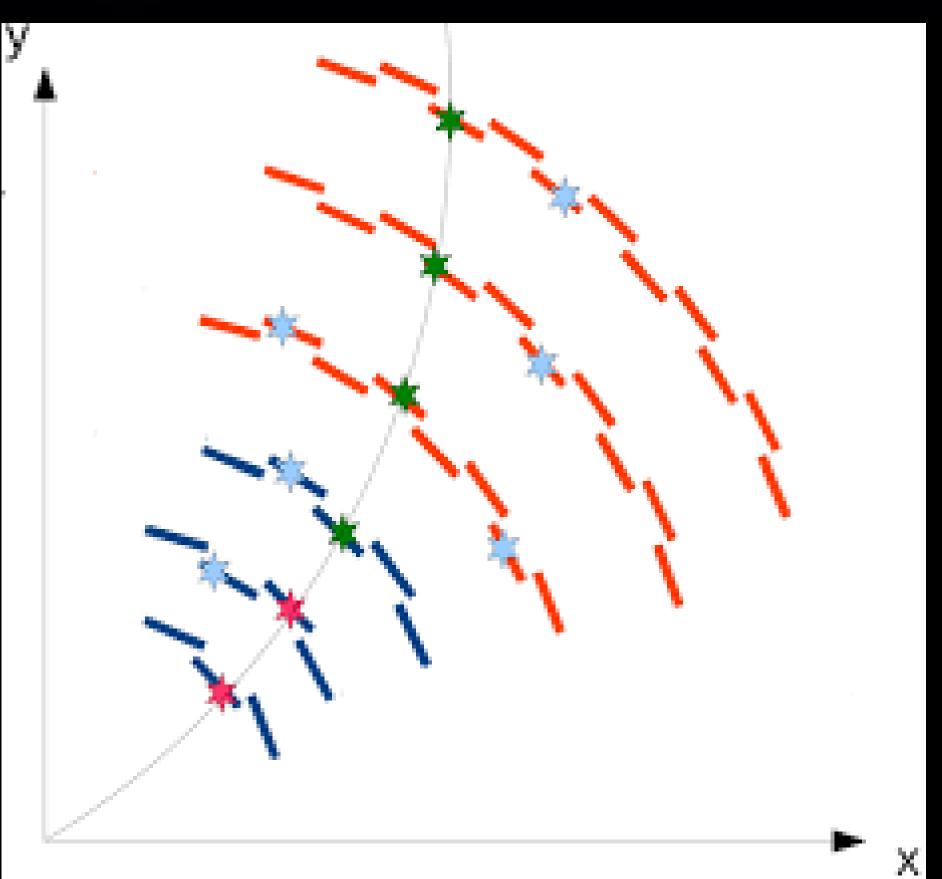
CERN: 250 GeV electrons passing through **27 X**₀.

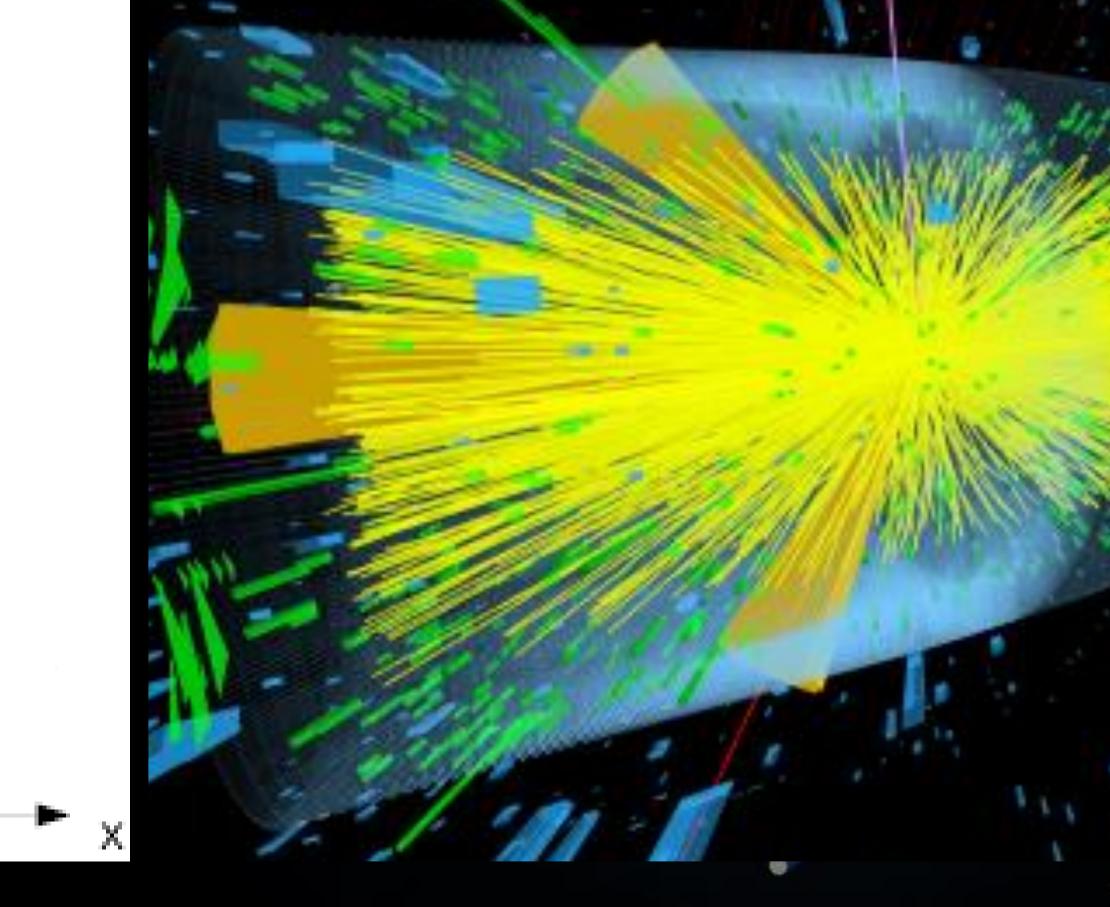


- You mean, apart from the small matter of 300Tb/s of data?
- So much data it has to be zero-suppressed
 - No (or, at least, limited) geometric timing which can be utilized
 - Variable data-volume
 - Do you handle the worst case? Very inefficient
 - Do you handle the average? How do you handle overflows?
- We did such a good job at Phase-I, people have very high expectations...



• Real-time track-finding and fitting







- Real-time track-finding and fitting
- Real-time vertex-finding





- Real-time track-finding and fitting
- Real-time vertex-finding
- 3D cluster-finding in endcap



muon gun @PU = 35

28+12+5 hits (gen-matched: 100%) χ²/ndf 34/90



- Real-time track-finding and fitting
- Real-time vertex-finding
- 3D cluster-finding in endcap
- Particle-flow

standalone mu
tracks
electromagne [.] calorimeter
hadronic calorim

