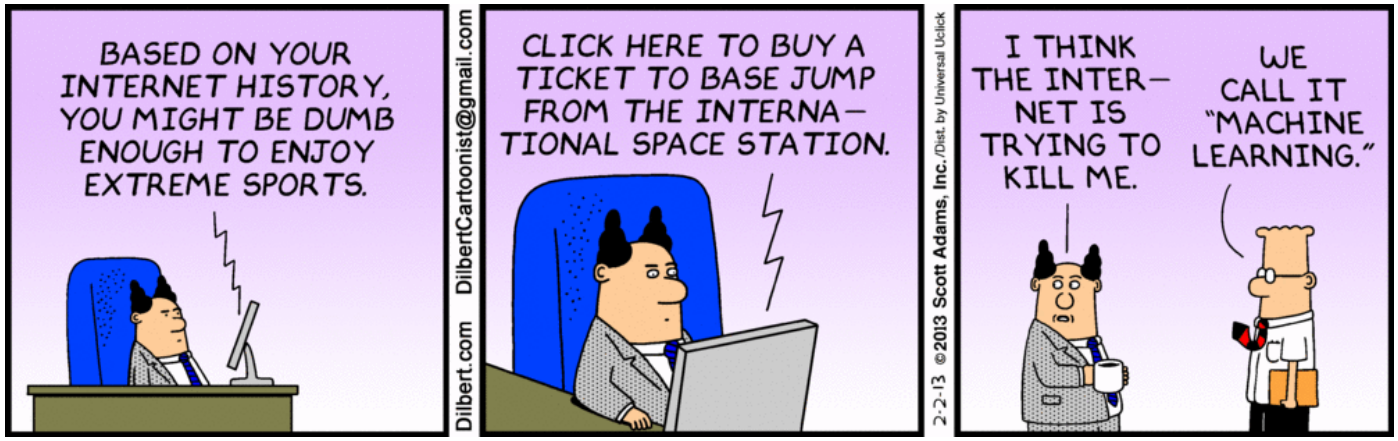
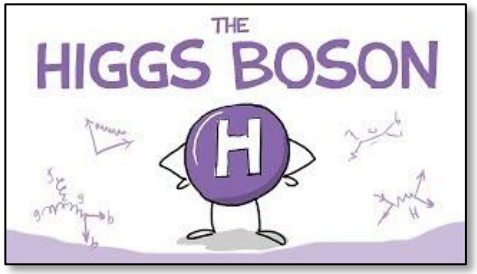
Réseaux de neurones embarqués: Capteurs CMOS Intégrant un RÉseau NEuronal

Auguste Besson
Projet CIRENE (mission interdisciplinaire 2016 CNRS)
Groupe PICSEL & plateforme C4PI
Christine Hu, Mathieu Goffe & Kimmo Jaaskelainen
Doctorant: Rui Guang Zhao (2015-2019)
Post-doctorant: Alejandro Perez (2014-2017)

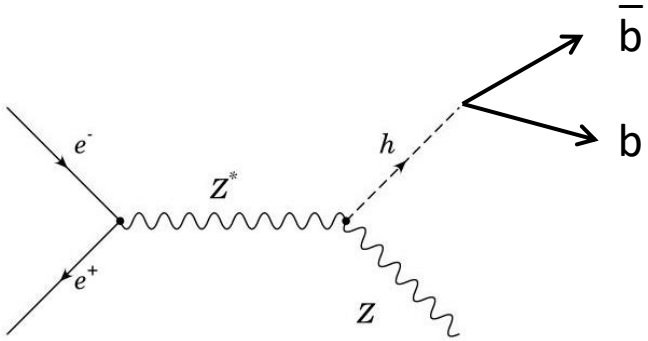


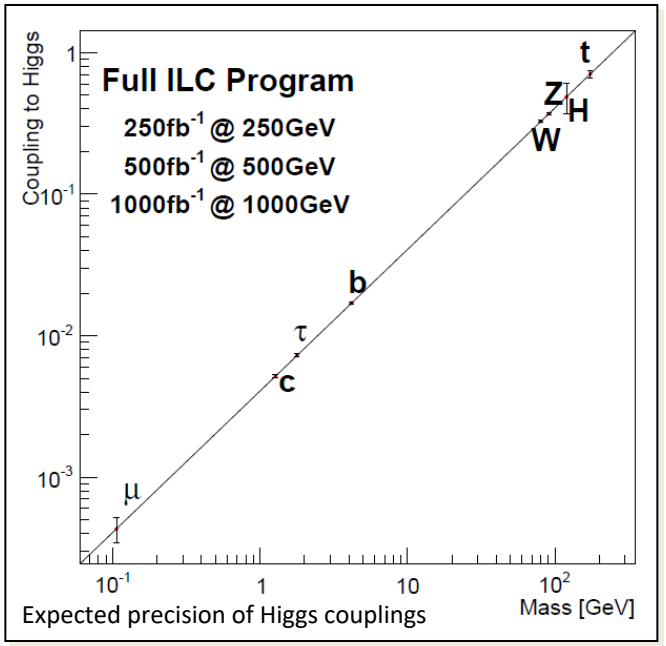
Les futures usines à Higgs (ILC)



- Le boson de Higgs a été découvert au LHC (2012)
  - ✓ Est-ce la fin de l'histoire ? NON !

- Il faut tester le Modèle Standard pour savoir ce qu'il y a « au-delà ». On peut:
  - ✓ Mesurer avec précisions certains paramètres
    - Par exemple la façon dont le boson de Higgs « se couple » avec les particules et leur conférer une masse.

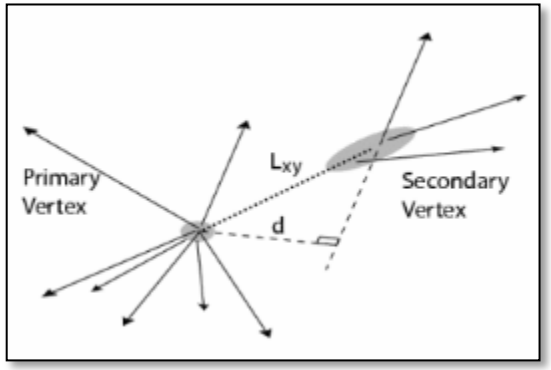


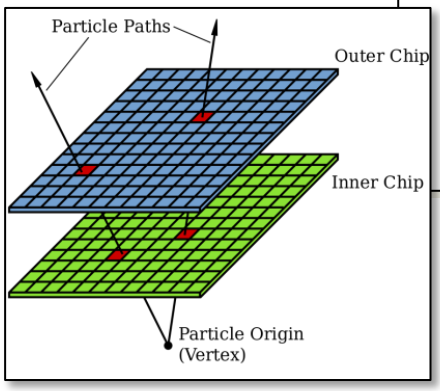


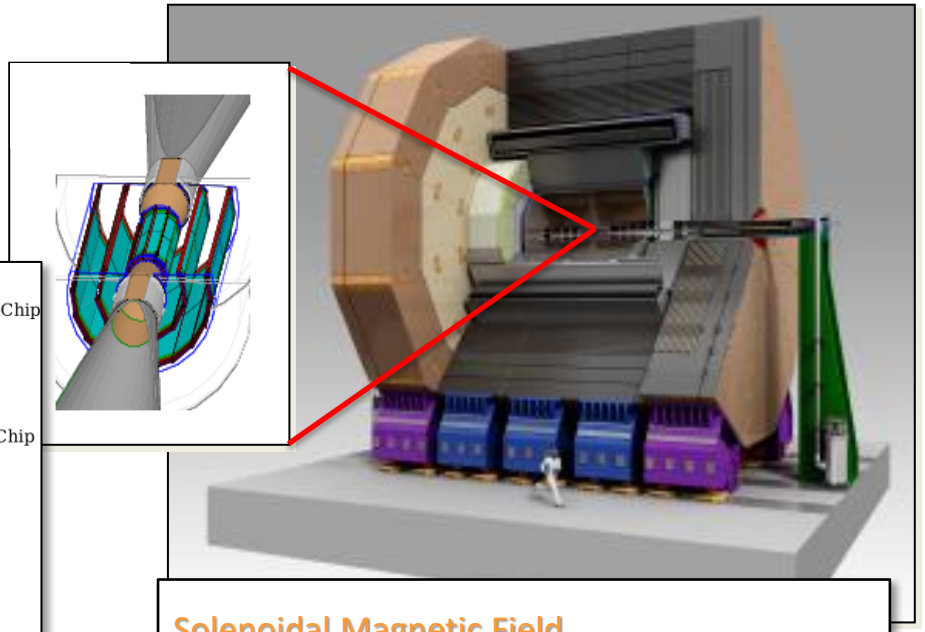
- Mise à jour de la stratégie européenne:
  - ✓ Un collisionneur e^+e^- (ILC, FCCee, CLIC, CEPC)
  - ✓ *«Higgs factory as the highest priority next collider»*

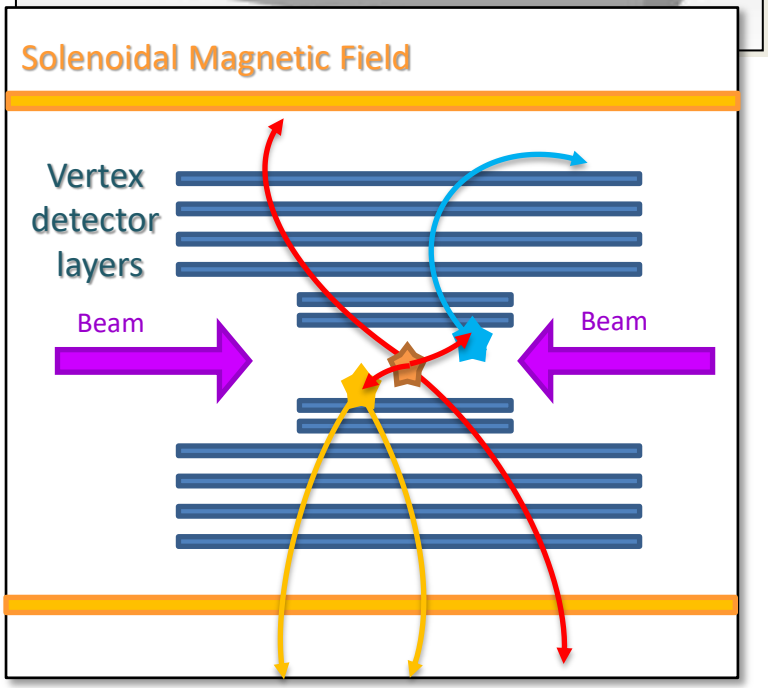
Le détecteur de vertex en physique des particules

- Reconstruire les vertex primaires et secondaires
  - ✓ Identification et reconstruction des processus: $e^+e^- \rightarrow Zh(Z \rightarrow \mu\mu)(h \rightarrow bb)$





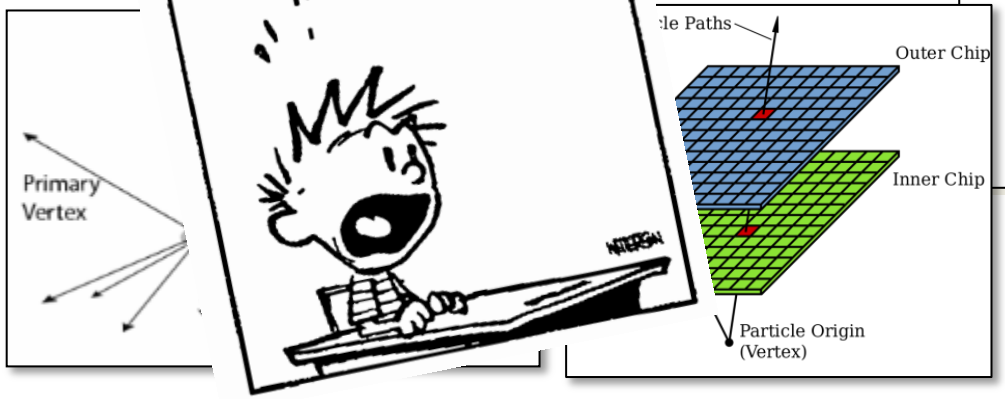


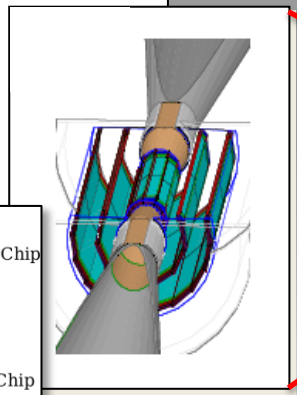


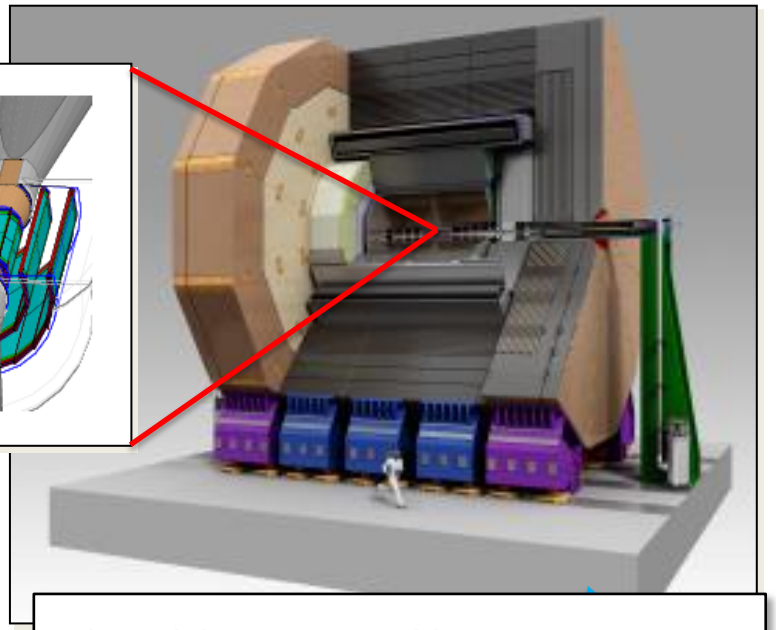
- Le détecteur doit être
  - ✓ Très trécis (~ 3 μm)
    - On veut reconstruire les vertex !
  - ✓ Le plus « transparent » possible
    - Sinon les particules sont déviées !
  - ✓ Assez rapide pour ne pas saturer
  - ✓ Ne pas dissiper trop de chaleur
    - Sinon il faut refroidir avec un système dédié et ce n'est pas transparent !

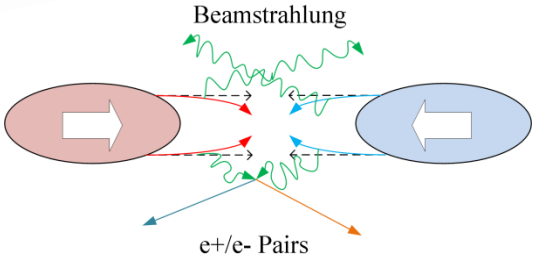
Le détecteur de vertex en physique des particules

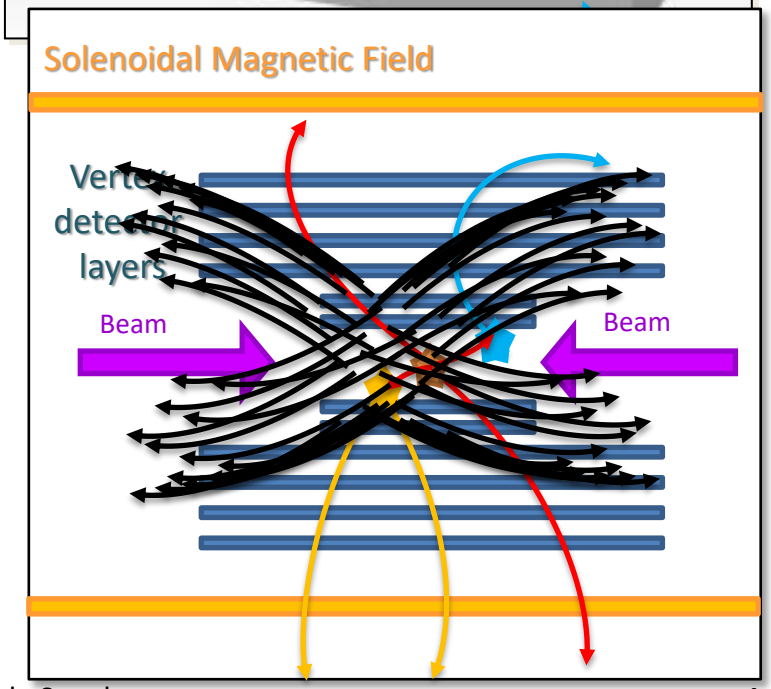
- Reco… …ex primaires et
  - ✓ …nstruction
  - de … Z→μμ) (h→bb)











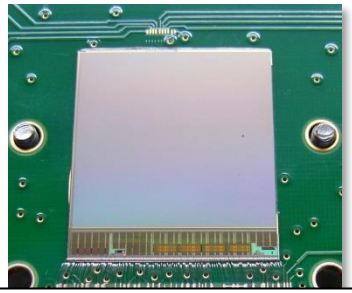
Les interaction entre les faisceaux produisent énormément de paires e^+e^- parasites de faible impulsion

⇨ Ces particules sont souvent « rasantes » (champ magnétique)

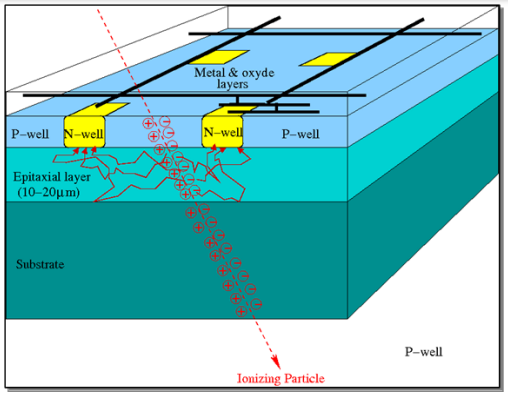
⇨ **Comment s'en débarrasser ?**

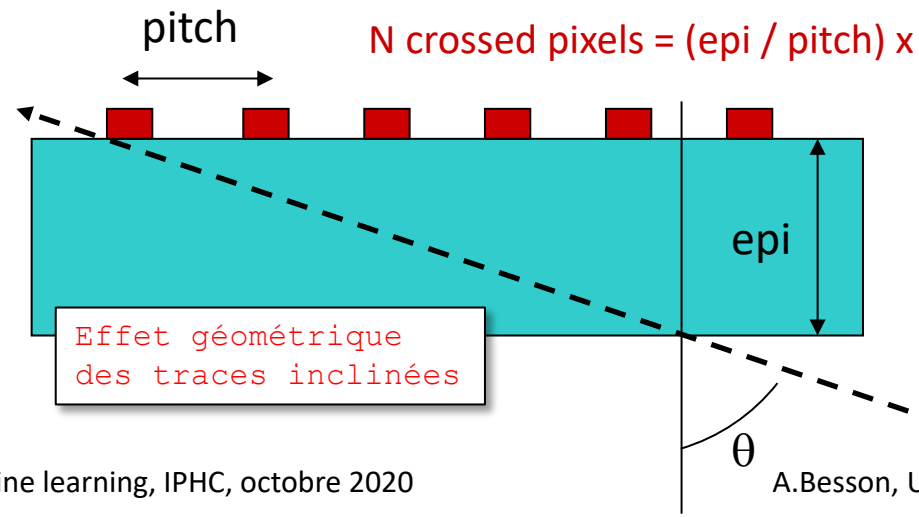
Capteur à pixels CMOS (PICSEL & C4PI)

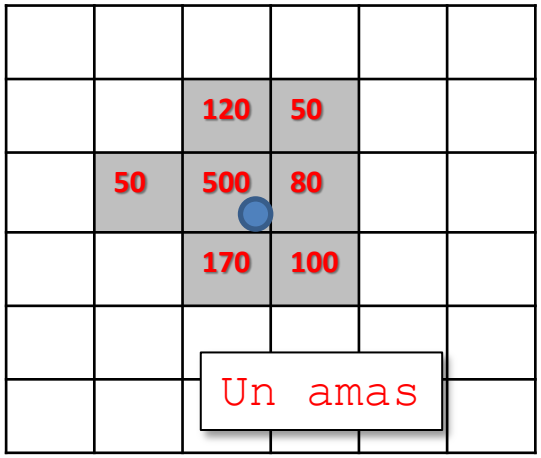
- Capteurs pixelisés
  - ✓ Passage d'une particule chargée
    - ionise la zone sensible
    - Diffusion des électrons
  - ✓ Charge collectée par plusieurs pixels
  - ✓ Ensemble des pixels collectant de la charge au dessus d'un seuil
    - = un amas (cluster)
    - Centre de gravité = position du passage de la trace



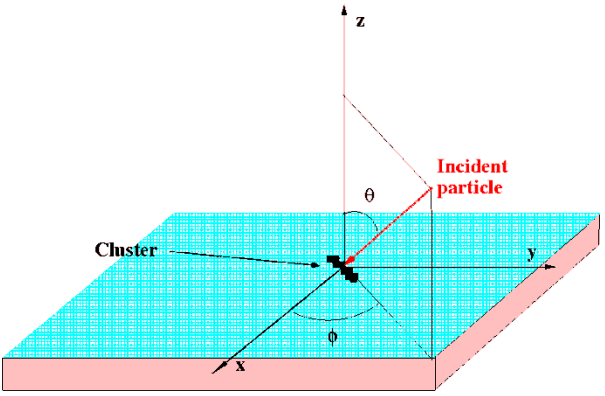
Un capteur développé à l'IPHC



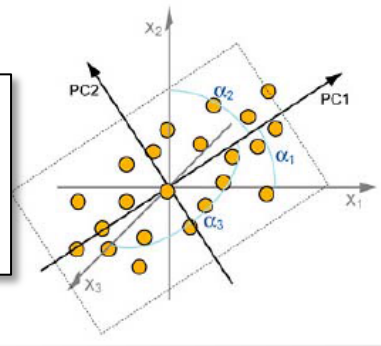


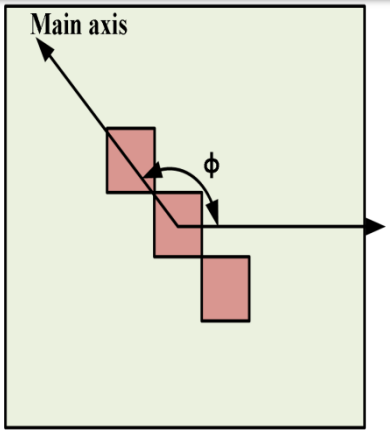


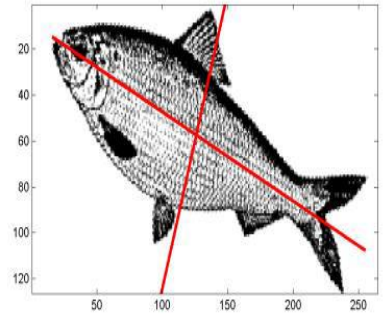
La forme de l'amas

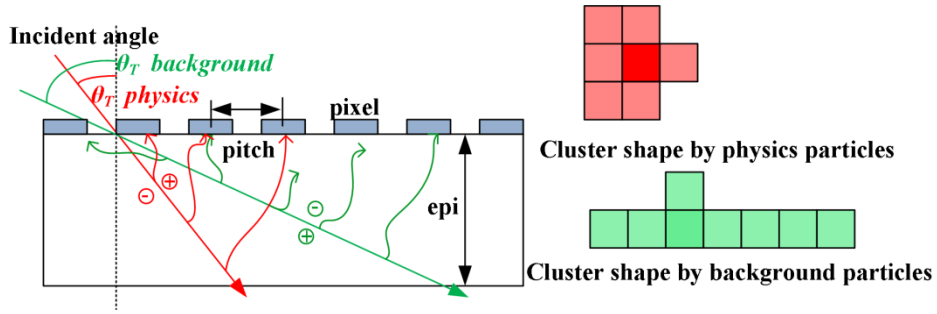


Analyse en composante principale
⇨ Axe principal
⇨ Reconstruire φ

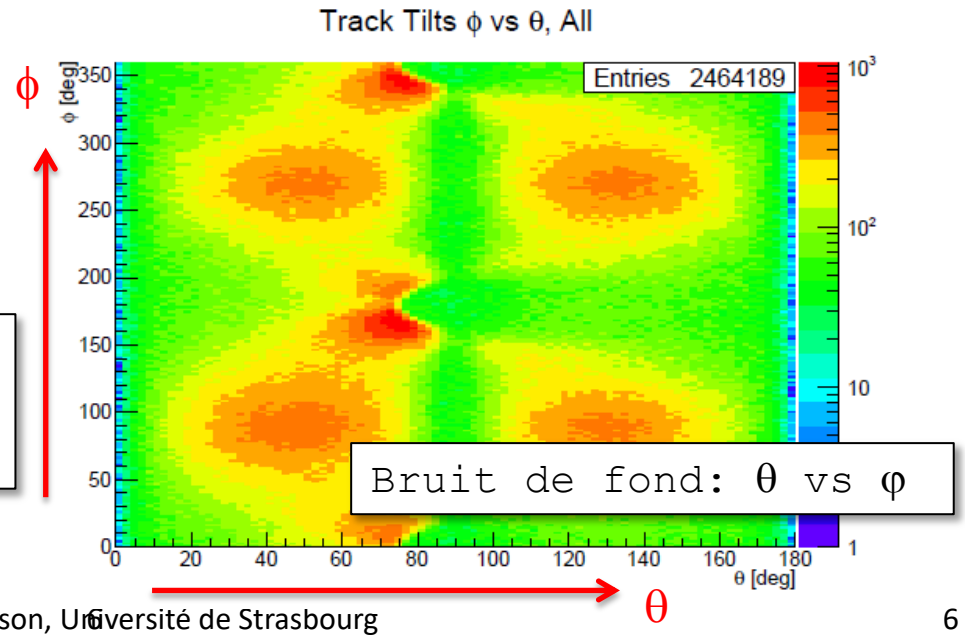








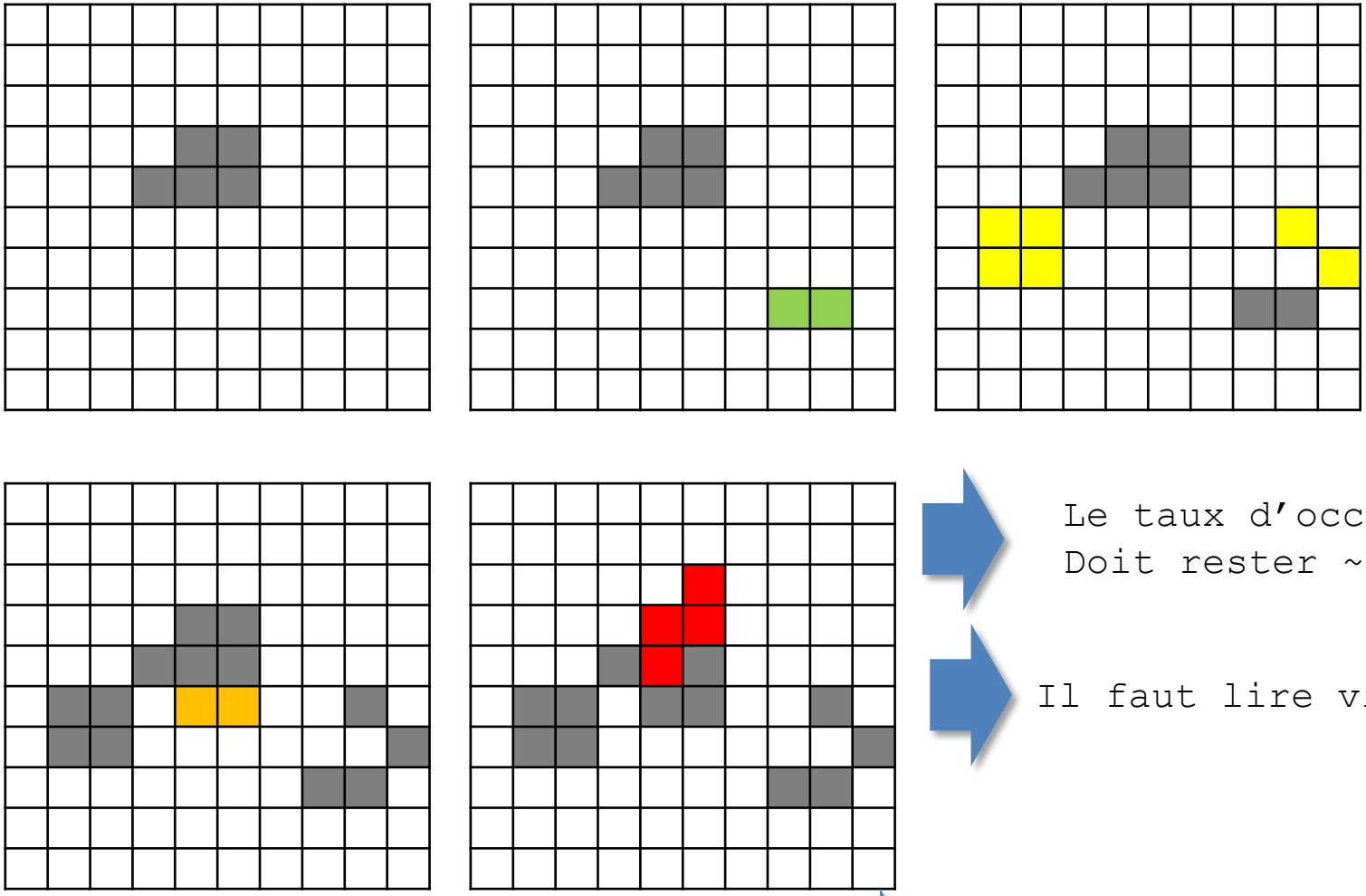
Élongation de l'amas
⇨ Reconstruire θ



Connaitre θ et φ
= identifier le bruit de fond

Bruit de fond: θ vs φ

Quand les impacts s'accumulent



Le taux d'occupation
Doit rester ~< 10^{-2}-10^{-3}

Il faut lire vite (~1 µs)

Le flot de données est énorme
~ 3 Gbits/sec (en moyenne)
~ 300 Gbits/sec (instantané)

Cahier des charges

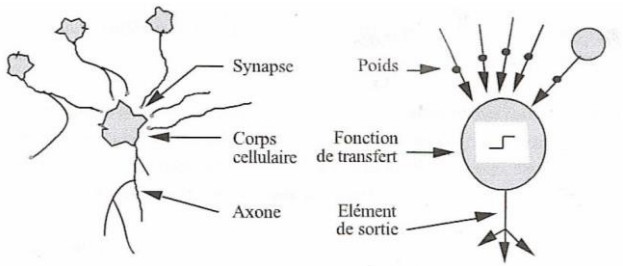
⇨ Utiliser un algorithme embarqué pour éliminer le bruit de fond sur la puce ?

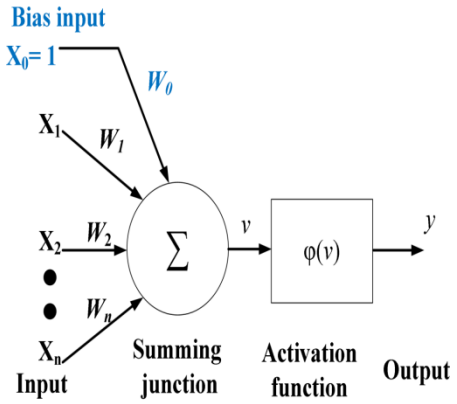
	Hors du détecteur	Sur une carte (FPGA)	Sur la puce
Précision	++	-	-
Filtrage / Extraction des données	-	+	++
Encombrement dans le détecteur	-	--	+
Puissance dans le détecteur	+	--	-
Stockage des données	-	+	++
configurabilité	+ +	+	-
coût	-	-	+

- L'algorithme doit reconstruire:
  - ✓ l'angle d'incidence θ et l'azymuth φ
  - ✓ Surtout pour les angles θ >~ 60°
- Algorithme suffisamment simple pour être potentiellement intégré sur une puce
  - ✓ Nombre réduit de paramètres d'entrée
  - ✓ Nombre réduit d'opérations
  - ⇨ Encombrement et Puissance limités
  - ⇨ Compromis sur les performances
- L'entrainement peut néanmoins se faire hors-ligne.

Les réseaux de neurones

- But: construire une fonction qui peut trier ou estimer une grandeur à partir d'un grand nombre de paramètres
- Structure de base: le perceptron

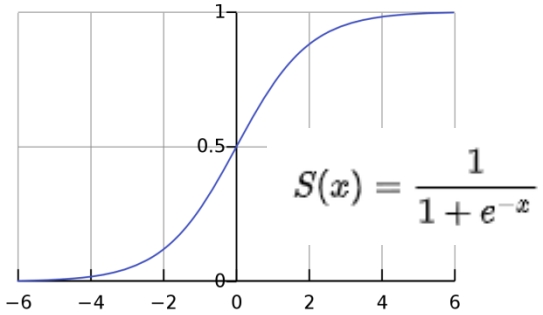


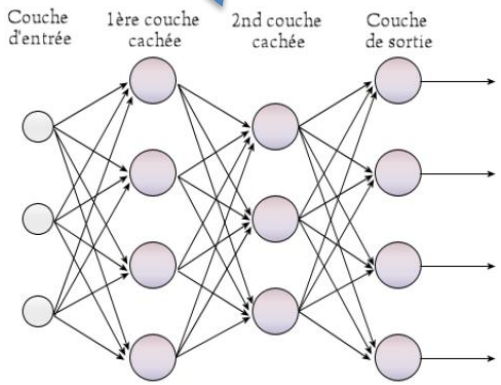


$$v = \sum_{i=1}^{n} x_i \times w_i + x_0 \times w_0$$

$$y = \varphi(v)$$

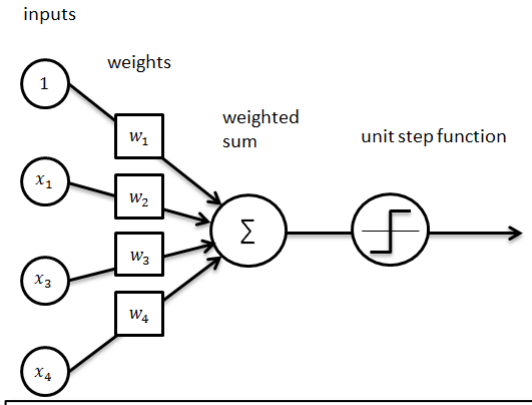
- Fonction de transfert
- Perceptron multicouche





- Entrainement:
  - ✓ détermination des poids w_i à l'aide de données pour lesquelles on connaît la « bonne » réponse
- Utilisation:
  - ✓ Une fois les poids connus, appliquer la fonction à n'importe quelle donnée

L'apprentissage



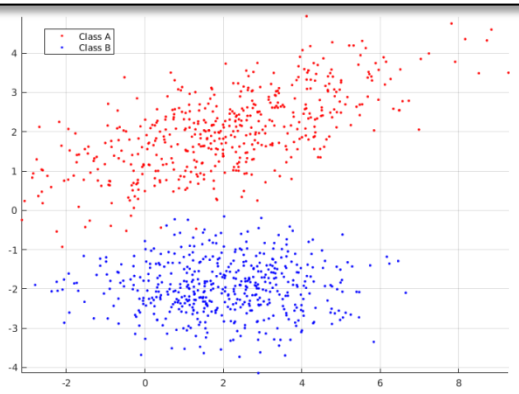
$$v = \sum_{i=1}^{n} x_i \times w_i + x_0 \times w_0$$

$$y = \varphi(v)$$

Apprentissage supervisé
(on sait ce qu'on cherche)
Par rétropropagation de l'erreur

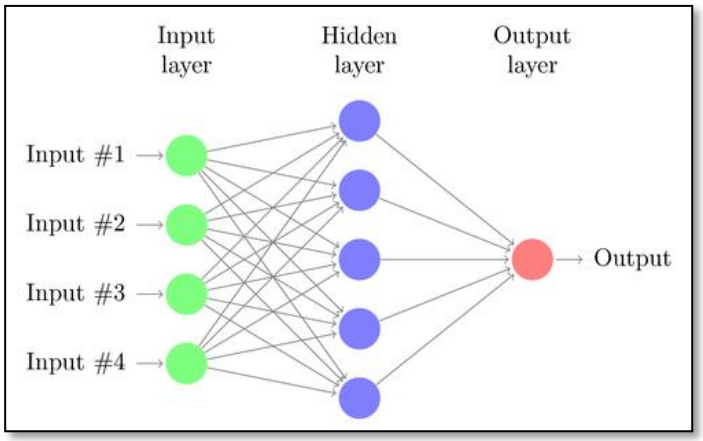
Si la réponse calculée est différente de la solution, on modifie les poids de manière à diminuer l´erreur commise par la cellule correspondante

- Problème de type régression
  - ✓ (≠ classification)
  - ✓ Solution = valeur continue entre 0° et 90°.
- Choix du jeu de données
  - ✓ Données simulées ⇨ on connait θ
  - ✓ Données réelles ⇨ on connait un peu moins bien θ
- Couvrir l'ensemble des angles possibles
  - ✓ Problème des bornes:
    - θ → 90° ⇨ amas infini
  - ✓ Problème des faibles angles
    - Entre 0° et 45° les amas sont très similaires
  - ✓ Numérisation
    - L'information de la charge est numérisée (ADC: 1-5 bits)

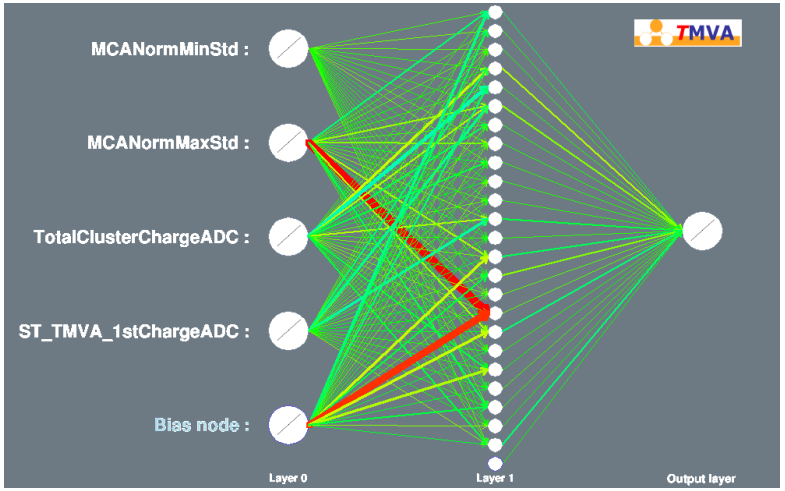


Choix de la structure du réseau de neurones

- Données d'entrée
  - ✓ Charge totale collectée
  - ✓ Charge du pixel siège
    - Ou charge de tous les pixels
  - ✓ Écart type de l'axe principal pondéré par les charges
  - ✓ Écart type du 2e axe

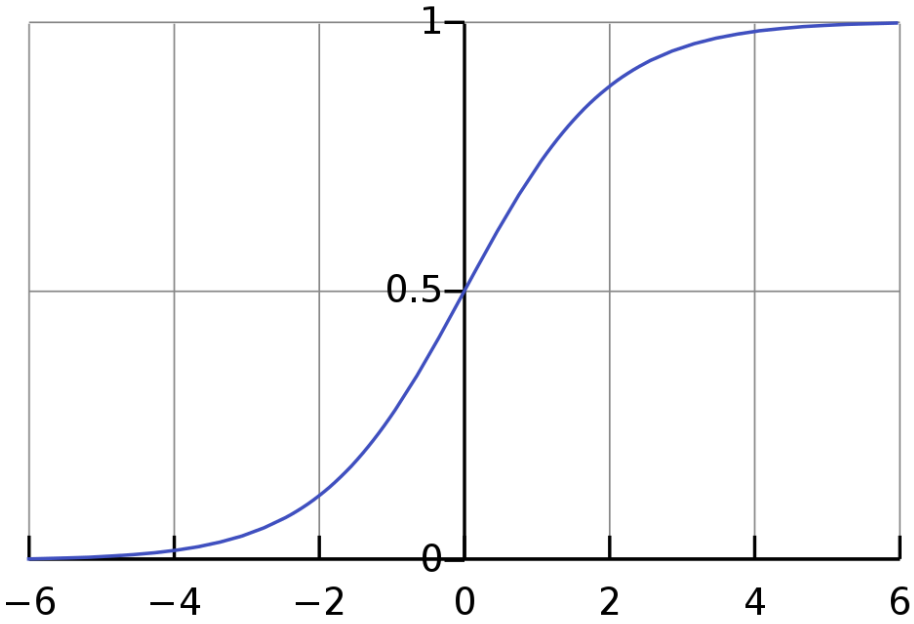


- Structure simple
  - ✓ 1 seule couche cachée (15 nœuds)
    - Si pas de couche cachée ⇨ uniquement des problèmes linéaires
  - ✓ 1 biais en entrée
- Sortie
  - ✓ Type régression
    - (≠ classification)
  - ✓ Une seule sortie continue
    - Estimation de θ



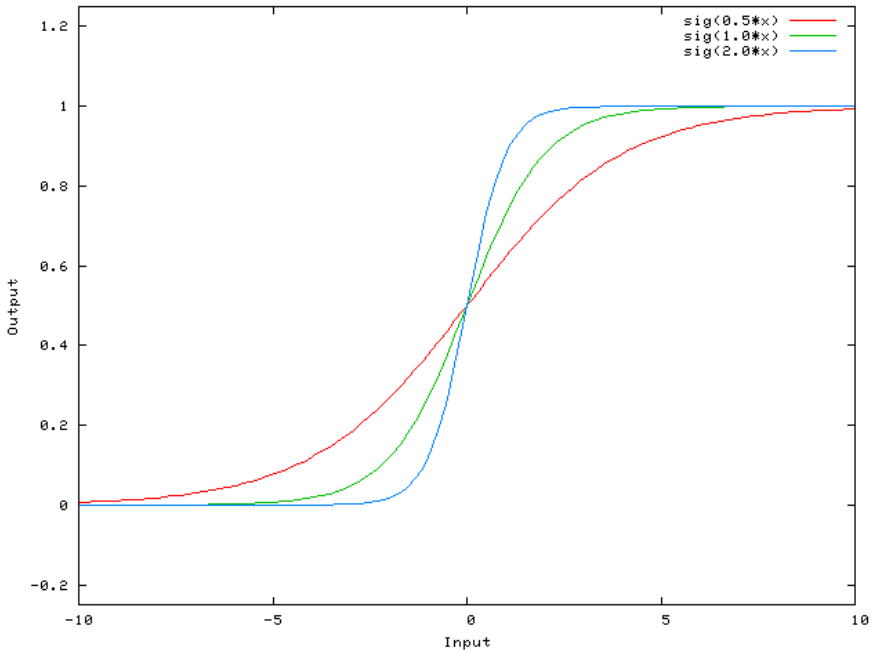
Fonction de transfert

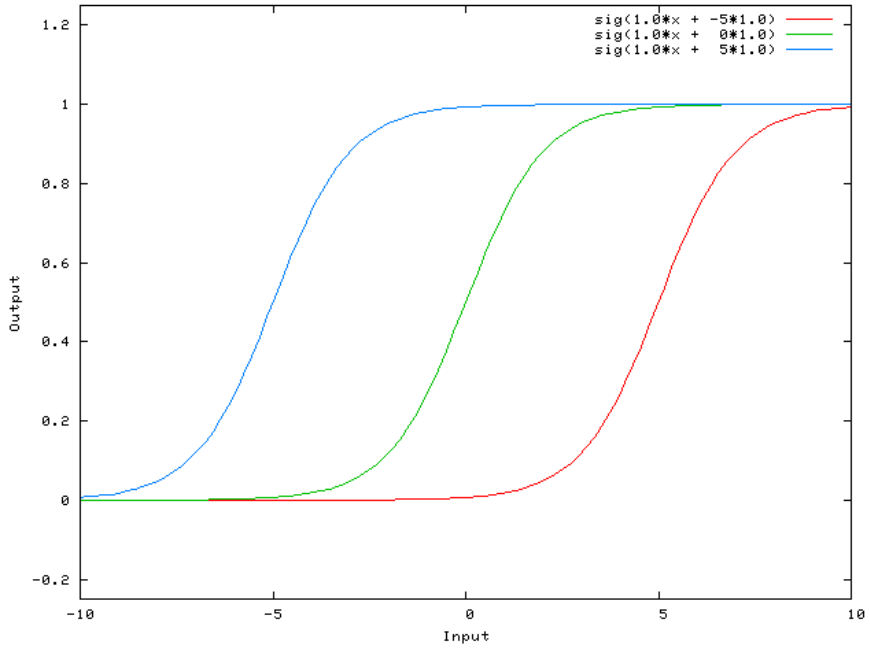
- Peut être n'importe quoi en principe
  - ✓ Sigmoides ,tanh, heaviside, etc.
  - ✓ Efficacité en général des fonctions bornées « à seuil »
  - ✓ Difficultés d'implémentation sur une puce
    - ⇨ fonctions approchées
    - « look up table » (tables de données)

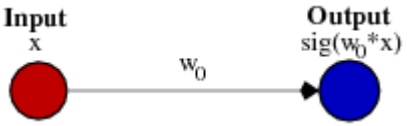


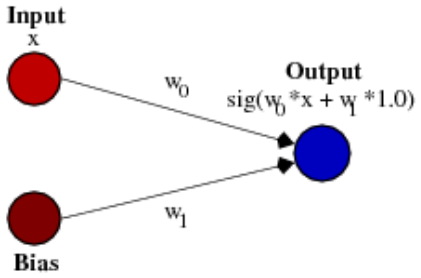
Rôle du biais

- **permet de décaler la fonction d'activation vers la gauche ou la droite**
  - ✓ critique pour un apprentissage réussi.

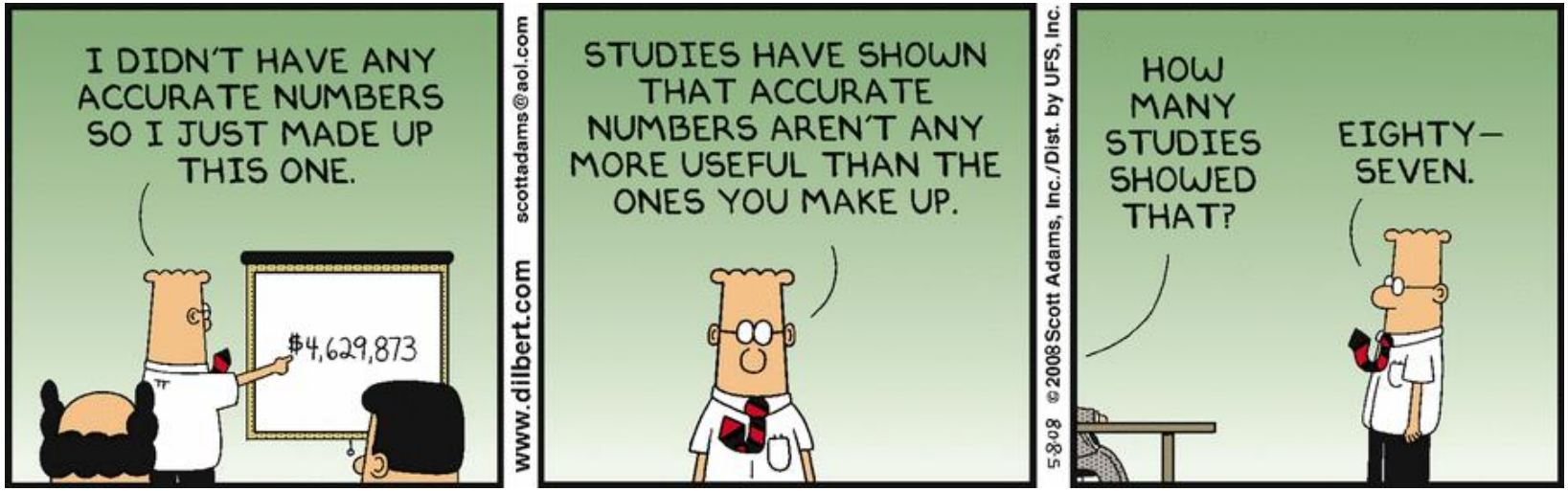




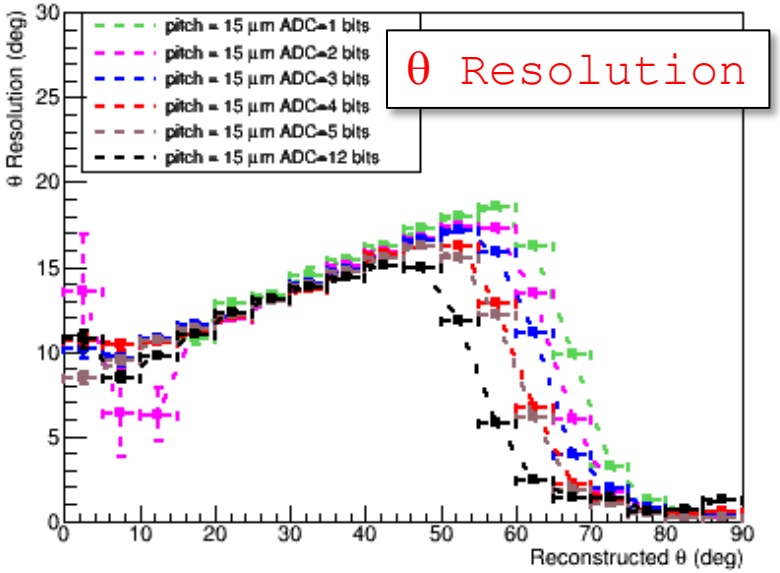


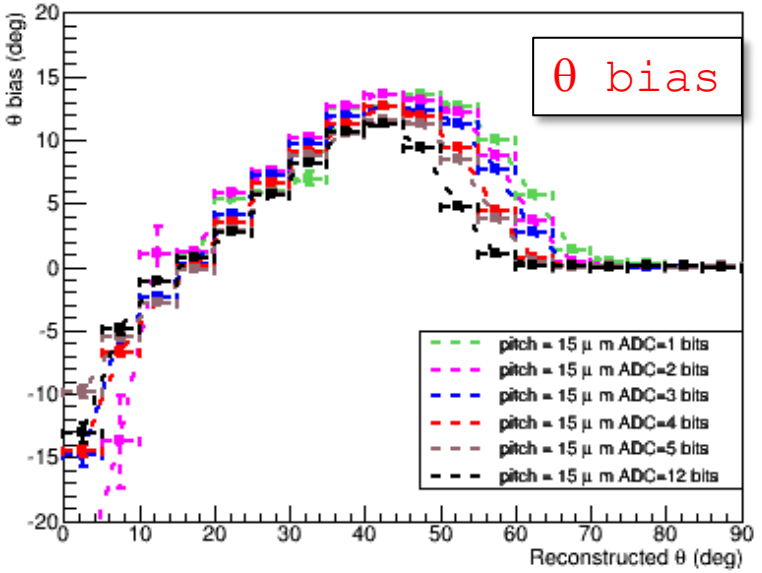


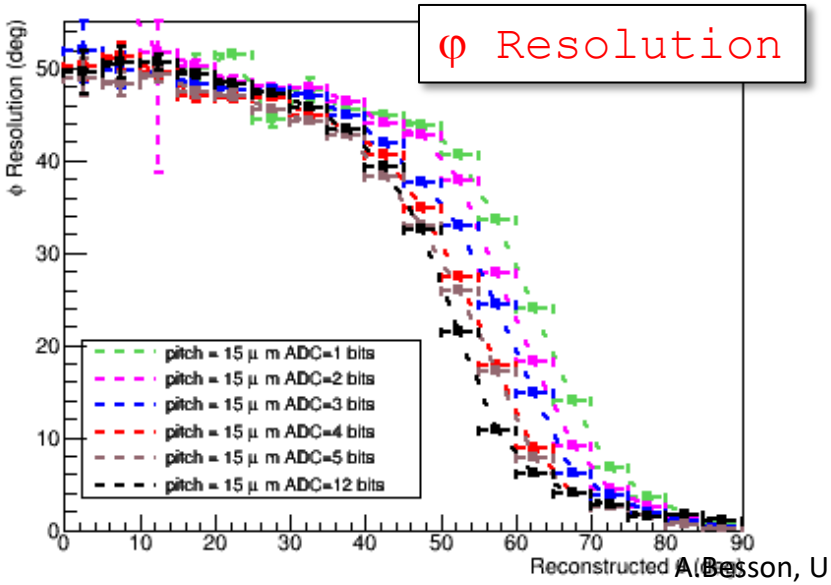
Quelques résultats

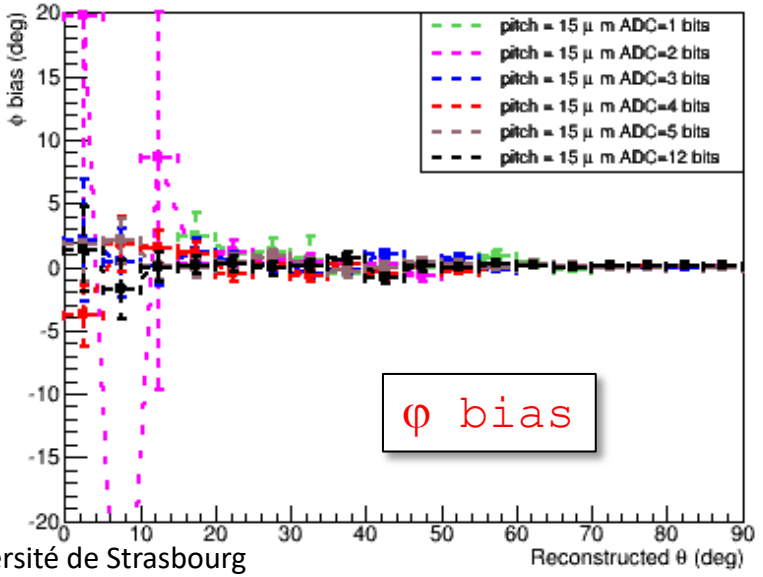


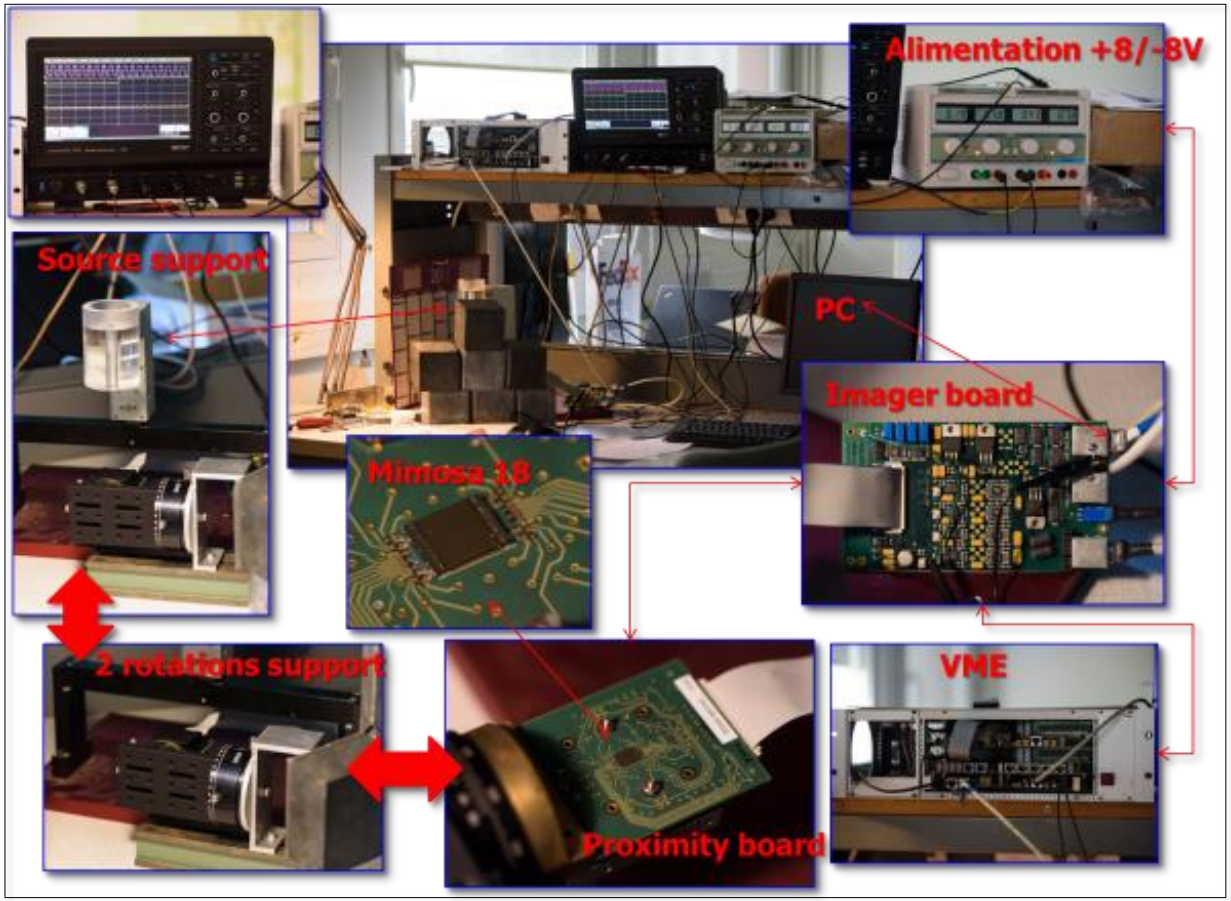
Performances en fonction de la numérisation

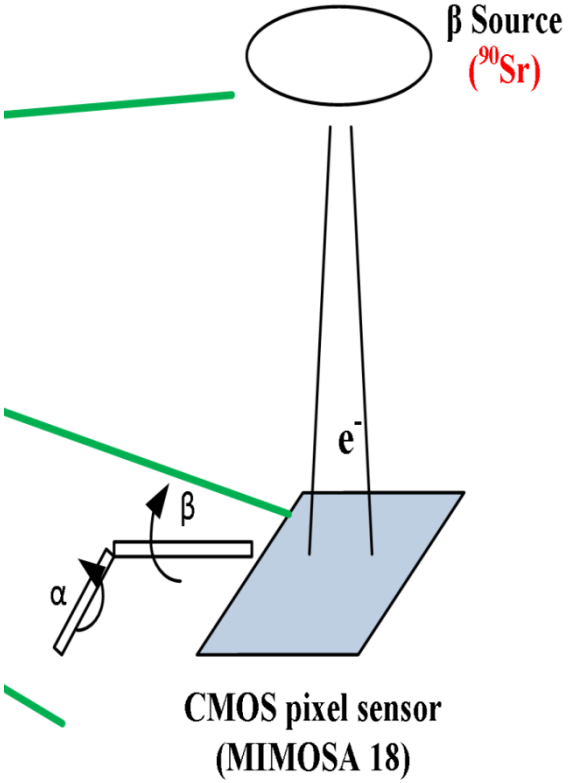










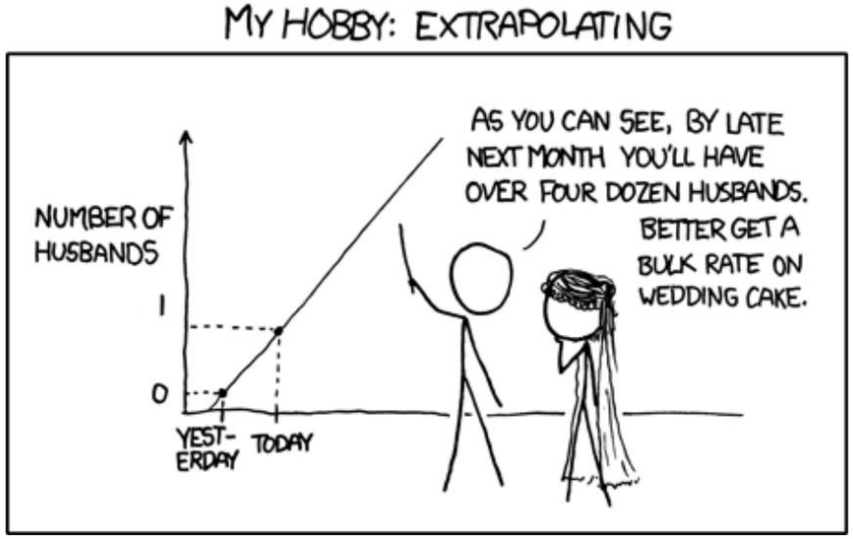


- A dark chamber (box)
- β^- source ^{90}Sr
- A CMOS pixel sensor (MIMOSA 18)
- A 2 rotations support

→ Performances Offline ~ FPGA
Confirme les données simulées

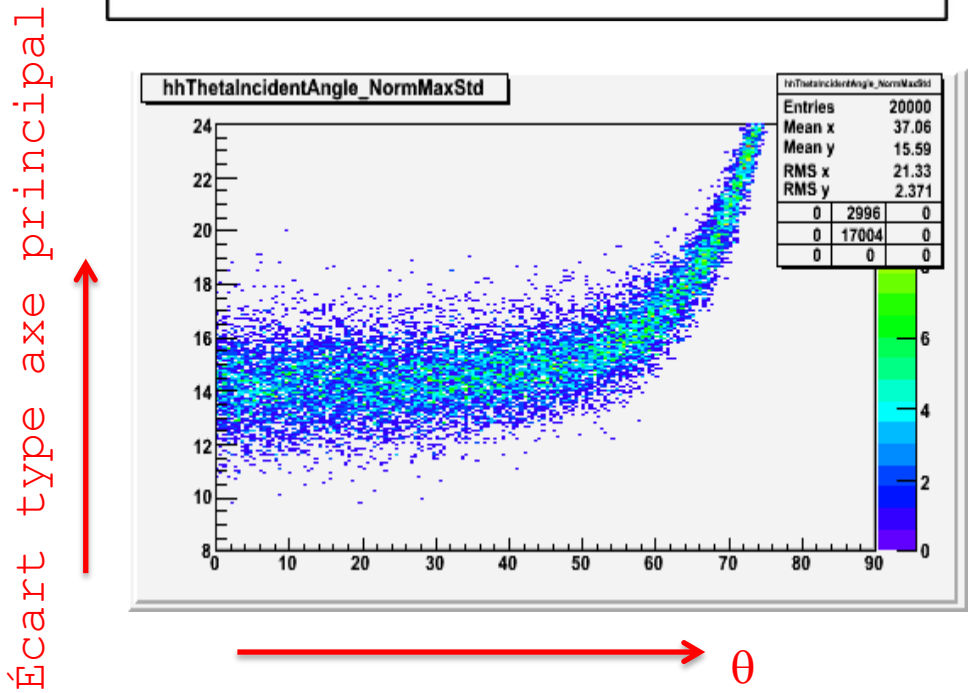
Apprentissage et extrapolation

- Le domaine de validité ne peut pas en principe être étendu en dehors du domaine où réside la population qui a servi à l'apprentissage



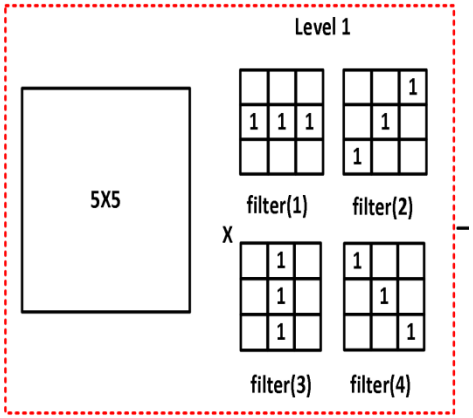
- Difficulté d'obtenir des données avec
  - ✓ θ → 90°
    - (taille des amas infinie)

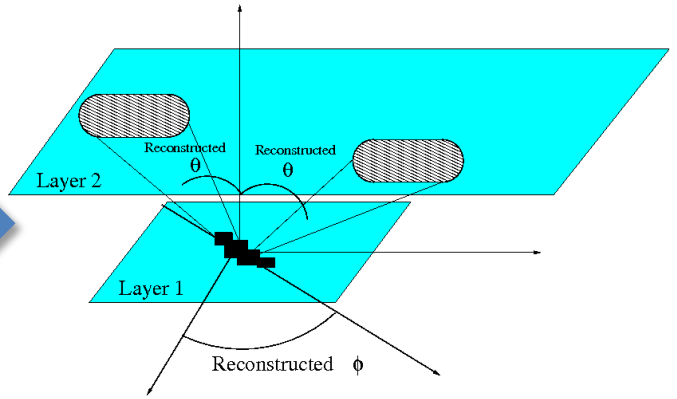
- Peu de sensibilité à petit angle

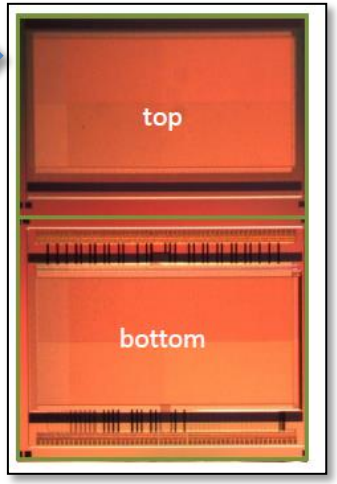


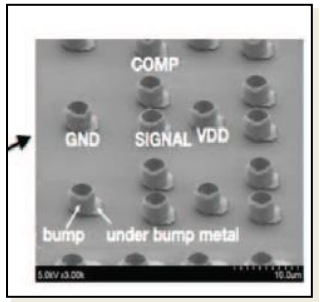
Perspectives

- Algorithmes complémentaires:
  - ✓ Mise en amas, Composantes principales sur le chip !
    - Problèmes en soi
    - Pre-Filtrage 3x3
  - ✓ Reconstruction des amas « 1 pass »
- Faisabilité ?
  - ✓ Limites techniques aujourd'hui
    - Puissance dissipée et complexité
  - ✓ Percée technologique ?
- Pistes à suivre
  - ✓ Combiner l'information de 2 couches
    - Angles plus précis
  - ✓ Mixage classification/régression ?
    - Déclencher le calcul que pour amas allongés
  - ✓ 65 nm réduction de taille de grille
    - Plus de transistors !
  - ✓ Capteurs double feuille
    - Optimiser une couche pour la partie numérique
- Tout système embarqué à terme
  - ✓ Le machine learning embarqué se généralise
    - En général dans des cartes dédiées (FPGAs)
    - Problématique spécifique: algos simples



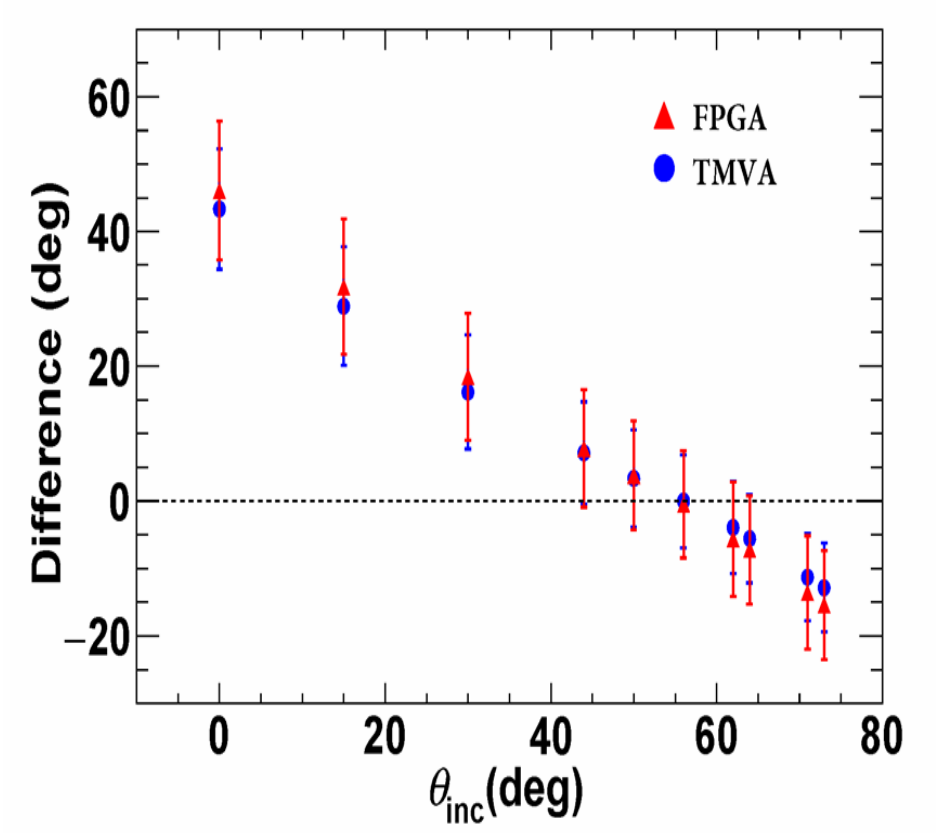




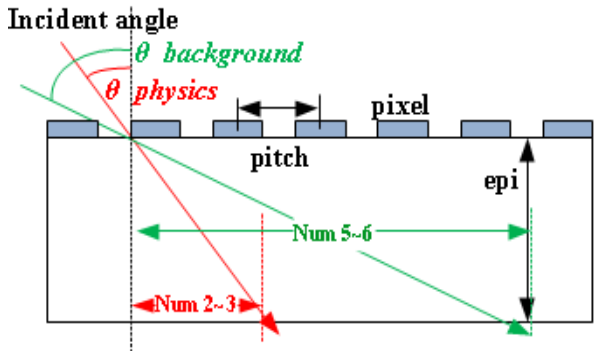


Dos haut

Données réelles: juste une preuve de principe



- **Angles reconstructed by two methods basically have the same mean value**



$$tan\theta = \frac{Num \times Pitch}{epi}$$

Minimum $\theta \approx 35$ degrees

- **Incident angle not well known (β source, collimation and multiple scattering issues)**
- **CMOS Prototype used not optimized for this application**
- **Incident angle < 35 degrees, cluster is composed by one pixel. It is not sensitive to cluster shape**
- **Smaller pitch, larger thickness epitaxial layer (by factory) ⇨ Minimum θ decrease**

Decorrelation: Principal Component Analysis

- PCA is typically used to:
  - reduce the dimensionality of a problem
  - find the most dominant features in your distribution by transforming
- The eigenvectors of the covariance matrix with the largest eigenvalues correspond to the dimensions that have the strongest correlation in the data set. Along these axis the variance is largest
  - sort the eigenvectors according to their eigenvalues
- Dataset is transformed in variable space along these eigenvectors

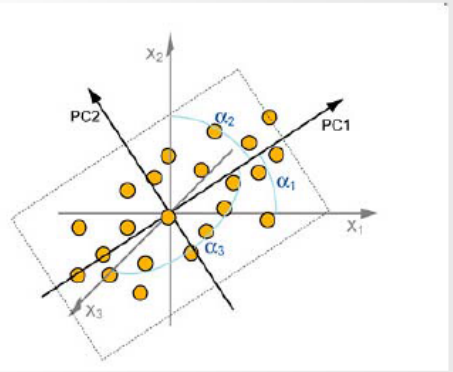
  → Along the "first" dimension the data show the largest "features", the smallest features are found in the "last" dimension.

$$x_k^{PC}(i_{event}) = \sum_{v \in \{variables\}} \left[x_v(i_{event}) - \bar{x}_v \right] \cdot v_v^{(k)} \quad , \forall k \in \{variables\}$$

Principle Component (PC) of variable k

sample means

eigenvector

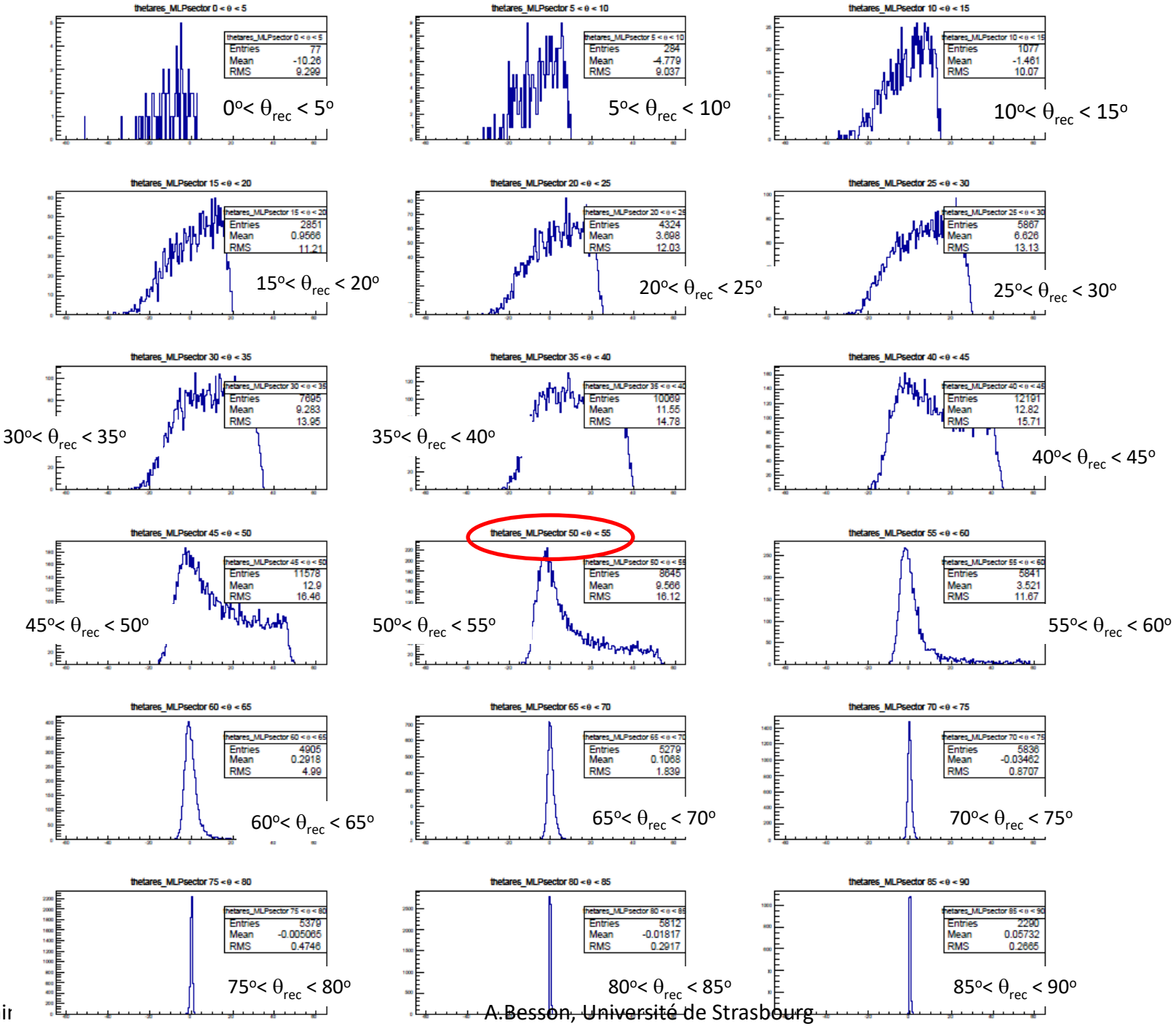


- Matrix of eigenvectors V obey the relation: $C \cdot V = D \cdot V$ → PCA eliminates correlations!

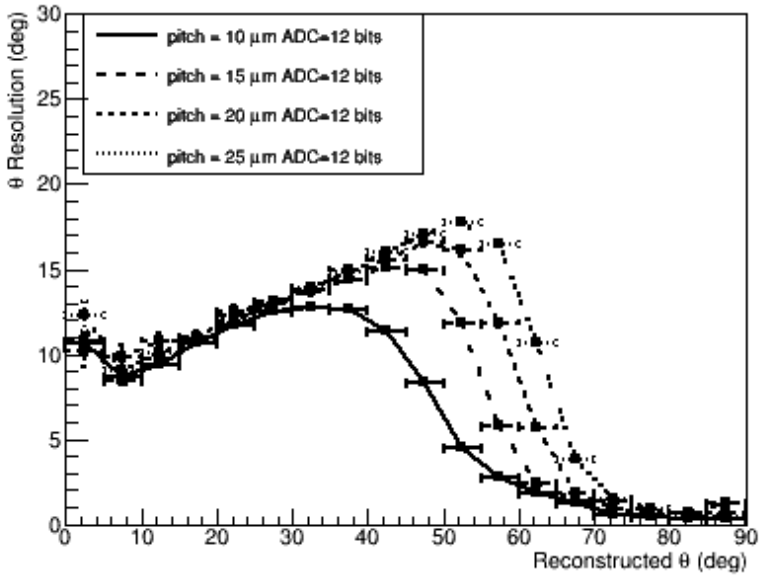
correlation matrix

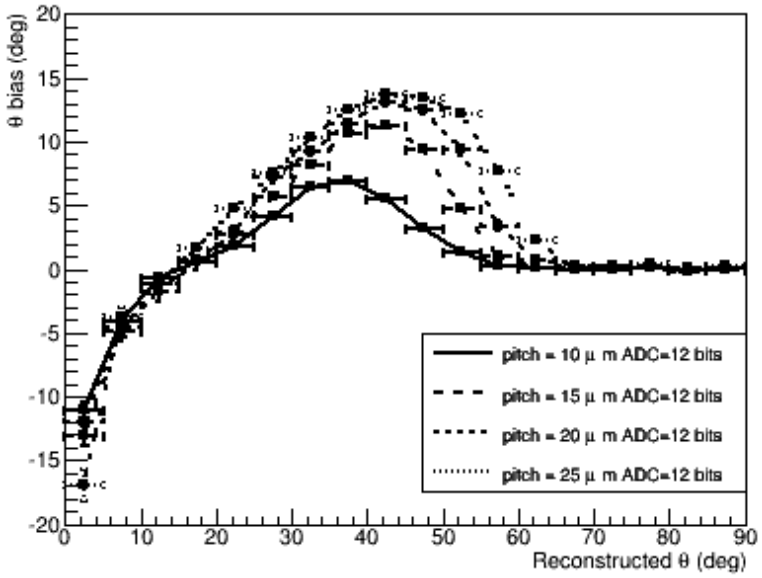
diagonalised square root of C

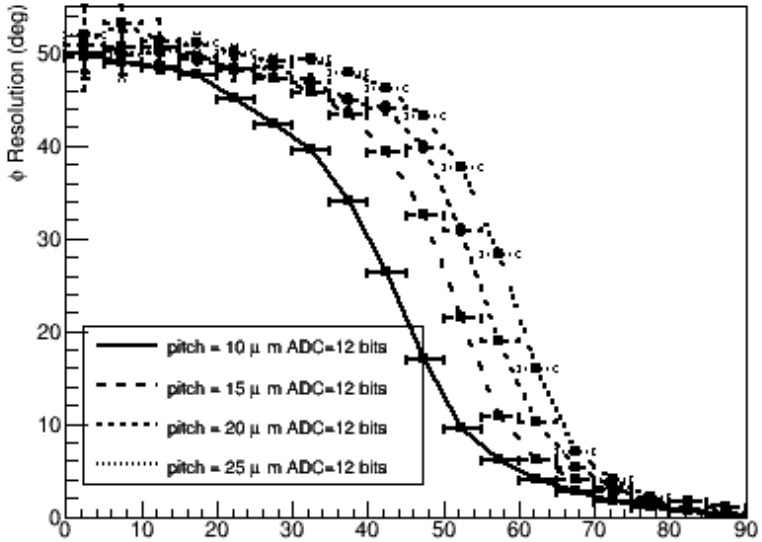
ADC 12 bits, pitch 20 um, epi 20 um

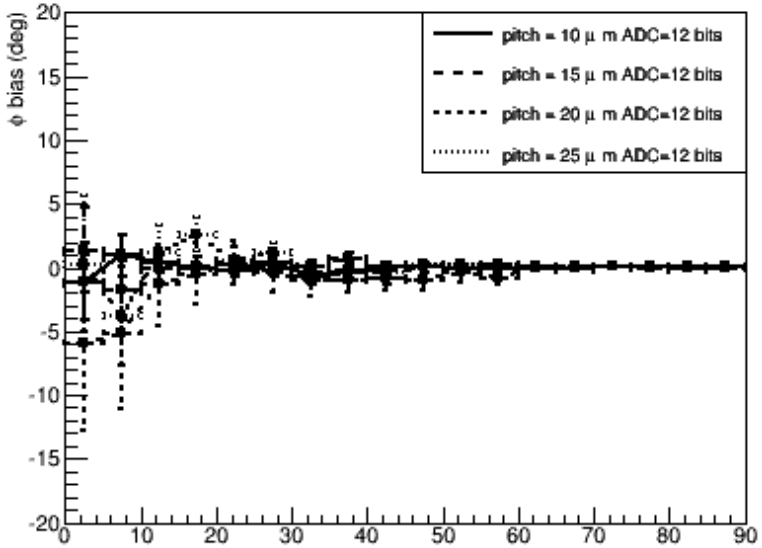


Performances vs pitch









64-column implementation synthesized using Cadence EDA TowerJazz 0.18 µm process *

Window	Multiplexer	ADC (bits)	Clock (MHz)	Column height (µm)	Column power (mW)
7×7	16-1	8	100	200	0.85
		8	200	200	1.83
		4	100	102	0.46
		4	200	103	0.88
7×7	32-1	8	100	197	0.86
		8	200	197	1.77
		4	100	101	0.43
		4	200	102	0.88
5×5	32-1	8	100	159	0.68
		8	200	159	1.45
		4	100	81	0.37
		4	200	82	0.71
7×7	64-1	8	100	196	0.87
		8	200	196	1.76
		4	100	101	0.43
		4	200	101	0.9

Occupied surface and power consumption can be reduced with

- **Small cluster window (High resistivity material of epitaxial layer)**
- **Low-frequency clock**
- **Low-resolution ADC**

No significant reduction on surface and power consumption for different multiplexer (Not taken MCA and ANN into account)

Optimization

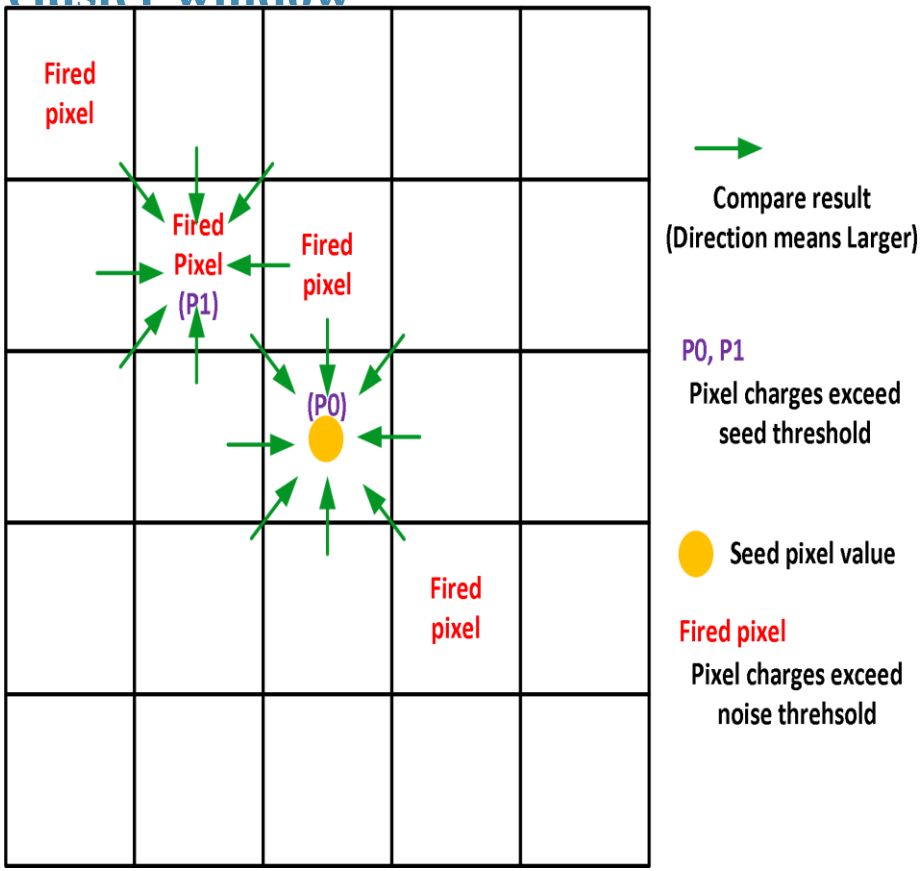
Power gating technology can be used to reduce power consumption

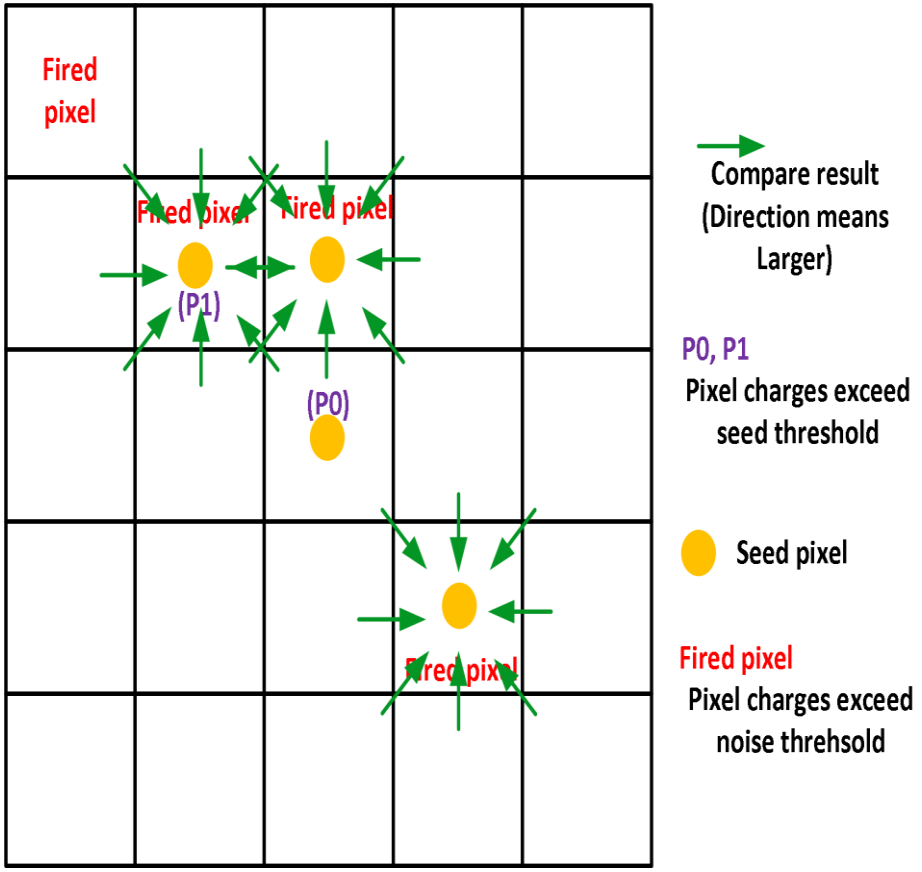
*

1, Column height is calculated according to the surface (pitch = 50 µm)

2, Multiplexer means the implementation of level 2

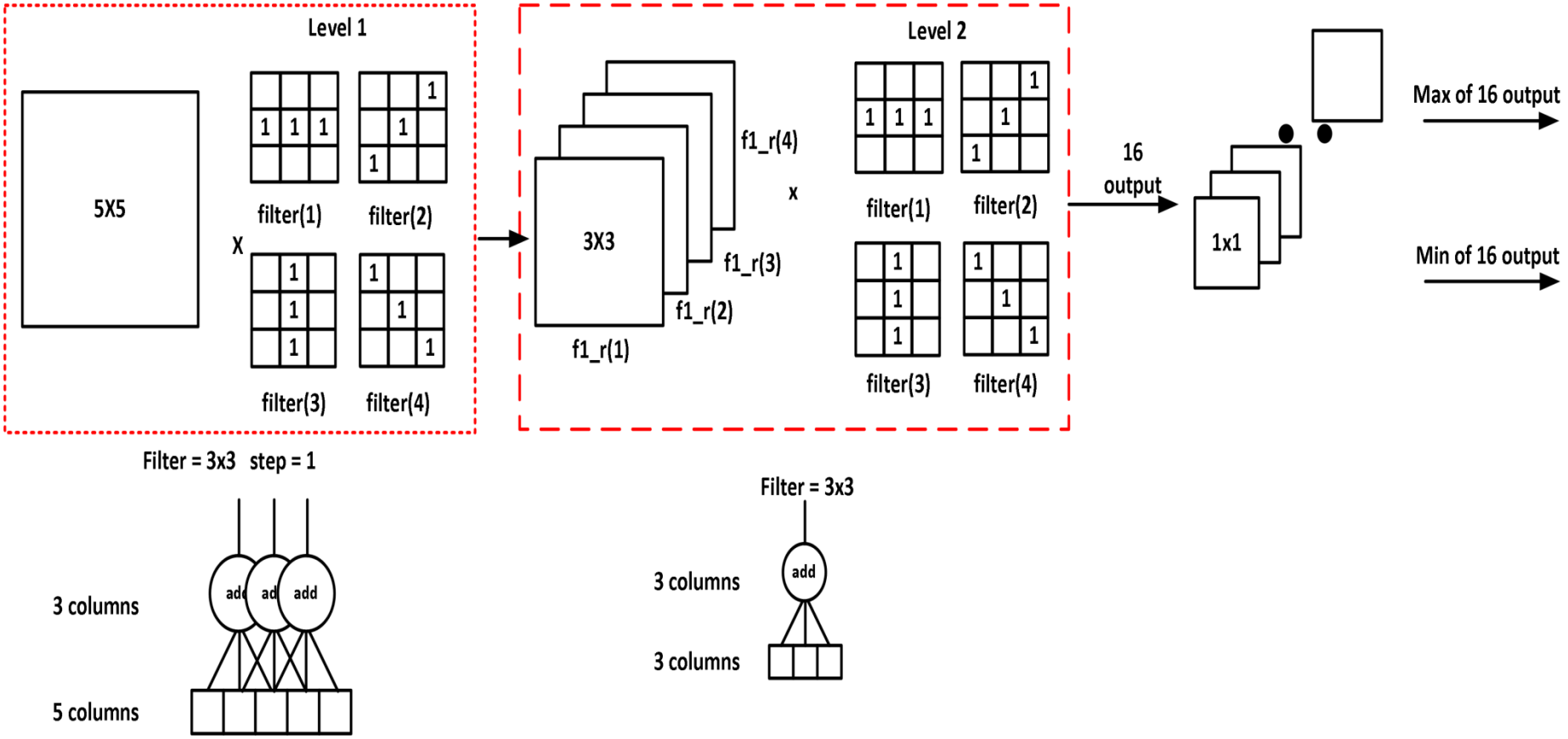
Algorithm with 5 × 5 cluster window, can be extended to 7 × 7 cluster window





Compared with 8 neighbor pixels, find seed pixels in a 3×3 cluster window

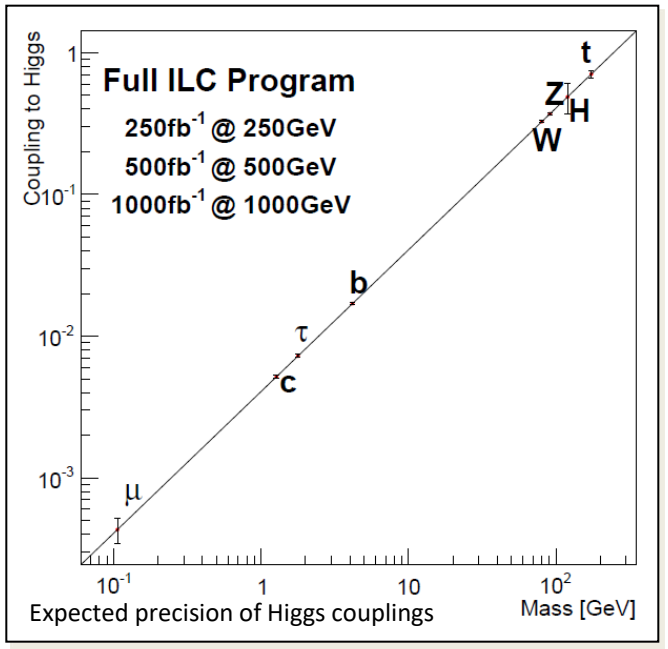
Replace fired pixels, repeat comparison, find seed pixel in a 5×5 cluster window



- Four operators are flowed on a cluster (5 × 5) at a step of 1 pixel and make convolution, respectively
- Four 3 × 3 submatrices are generated, then convolution processes are repeated on these submatrices
- The maximum and minimum value can be picked out from 16 values to present features of a cluster

ILC physics goals

- Very rich and wide program:
  - ✓ Top physics
  - ✓ EW precision physics
  - ✓ Direct/indirect searches BSM
    - DM search, exotic searches, etc.
- Higgs physics
  - ⇨Higgs parameters precise measurements
    - Mass, couplings, spin, etc.
    - Up to the % level
  - ✓ Model independent
  - ➢ Access to σ AND σxBr
  - ➢ Probe to new physics, model identification



Energy	Reaction	Physics Goal
91 GeV	$e^+e^- \to Z$	ultra-precision electroweak
160 GeV	$e^+e^- \to WW$	ultra-precision W mass
250 GeV	$e^+e^- \to Zh$	precision Higgs couplings
350–400 GeV	$e^+e^- \to t\bar{t}$	top quark mass and couplings
	$e^+e^- \to WW$	precision W couplings
	$e^+e^- \to \nu\bar{\nu}h$	precision Higgs couplings
500 GeV	$e^+e^- \to f\bar{f}$	precision search for Z'
	$e^+e^- \to t\bar{t}h$	Higgs coupling to top
	$e^+e^- \to Zhh$	Higgs self-coupling
	$e^+e^- \to \tilde{\chi}\tilde{\chi}$	search for supersymmetry
	$e^+e^- \to AH, H^+H^-$	search for extended Higgs states
700–1000 GeV	$e^+e^- \to \nu\bar{\nu}hh$	Higgs self-coupling
	$e^+e^- \to \nu\bar{\nu}VV$	composite Higgs sector
	$e^+e^- \to \nu\bar{\nu}t\bar{t}$	composite Higgs and top
	$e^+e^- \to \tilde{t}\tilde{t}^*$	search for supersymmetry

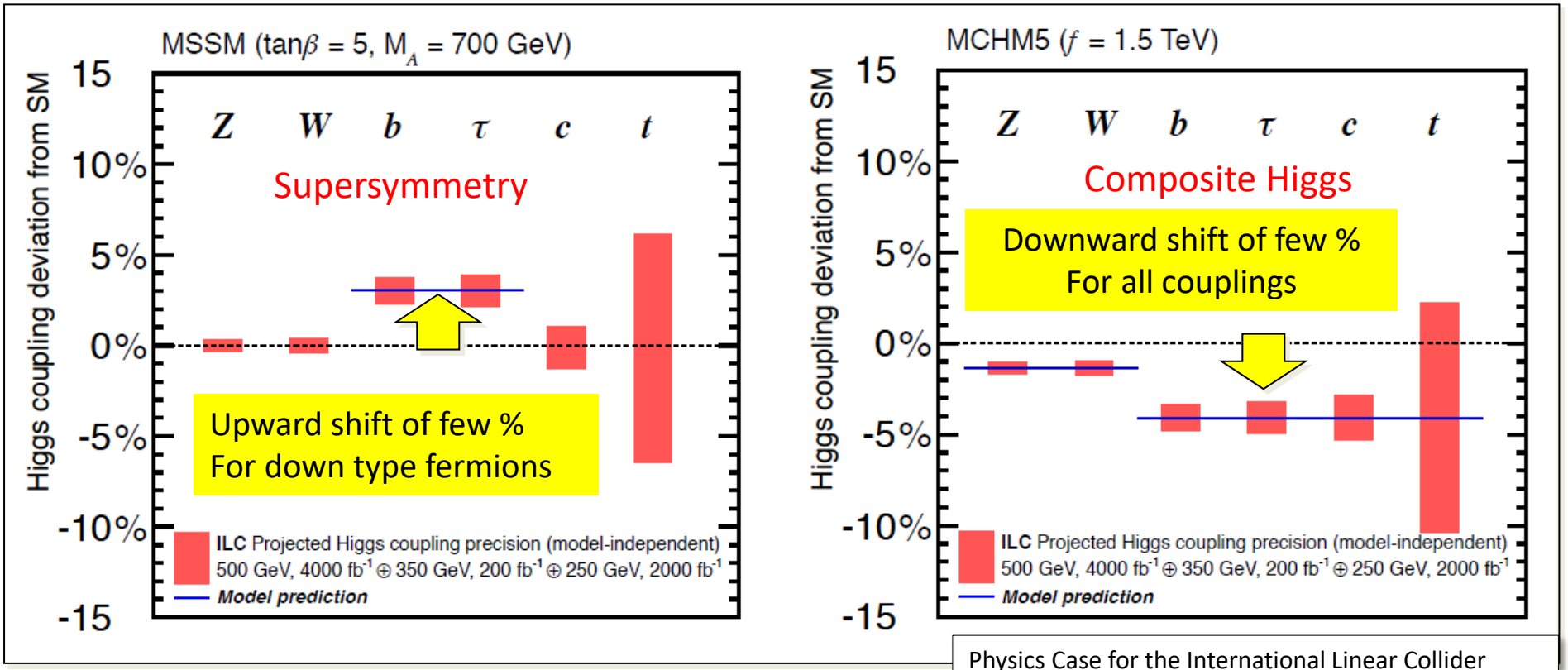
- Vertex detector role:
  - **b,c,τ tagging everywhere**
  - Low momentum tracking
  - Jet charge determination

Higgs boson couplings shifts in BSM (examples)

- Is the % level on the coupling precision enough ?
  - ✓ Size of deviation depends on new physics scale

$$\frac{g_{hbb}}{g_{h_{SM}bb}} = \frac{g_{h\tau\tau}}{g_{h_{SM}\tau\tau}} \simeq 1 + 1.7\% \left(\frac{1\ \text{TeV}}{m_A}\right)^2$$

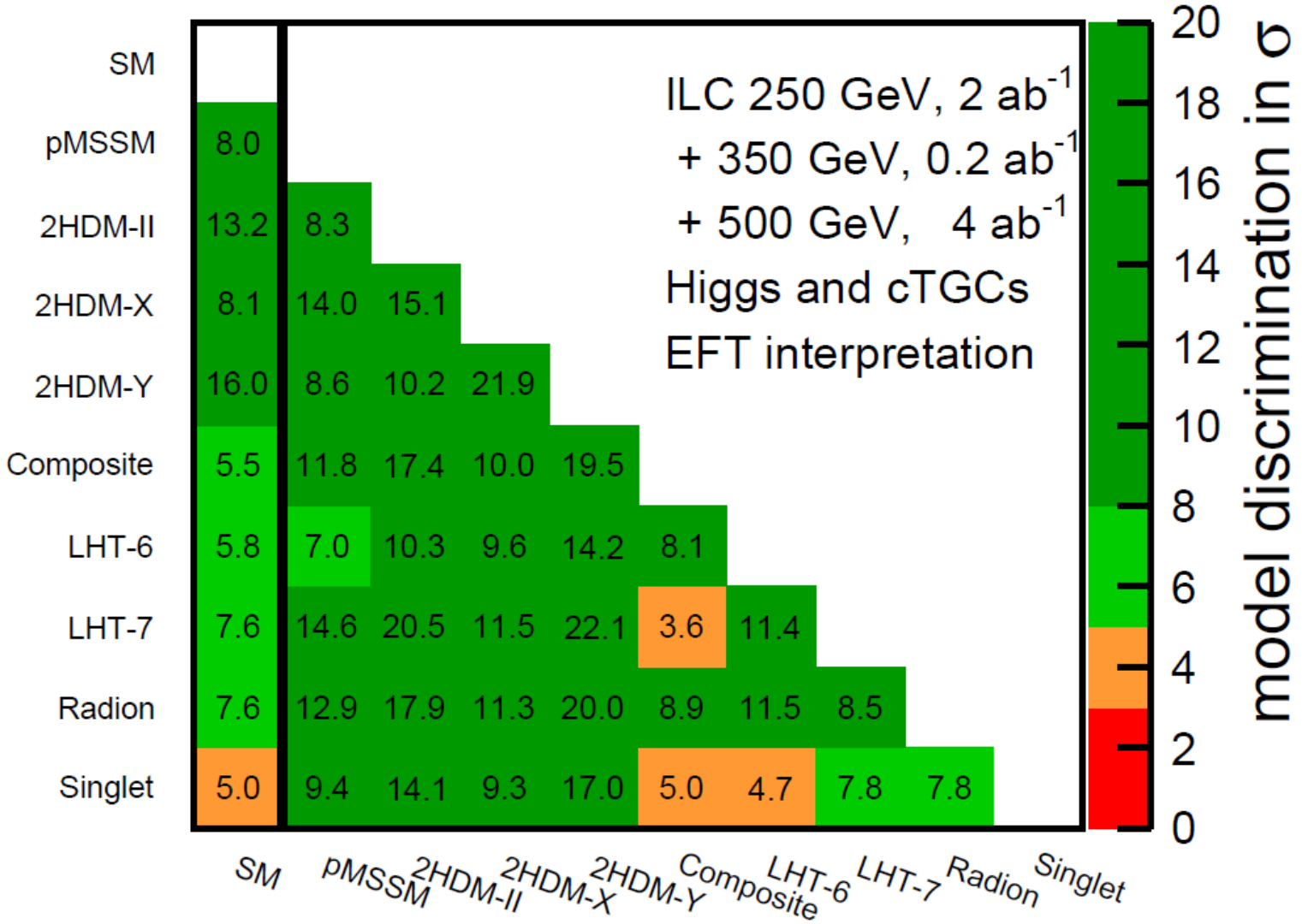
heavy Higgs mass



Physics Case for the International Linear Collider
K. Fuji et al. ILC-NOTE-2015-067

= 1 σ expected uncertainties from the full ILC data set (model-independent fit)

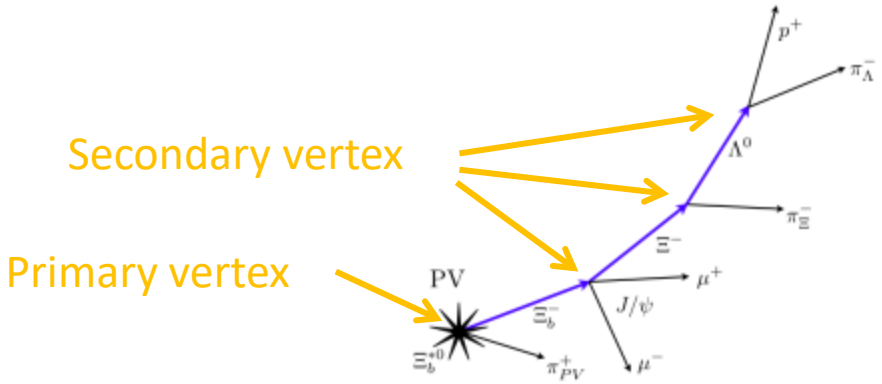
Model discrimination with ILC full data set



Graphical representation of the χ^2 separation of the Standard Model

Why do we need vertexing ?

- Reconstruct vertex to reconstruct the decay chain



- Heavy flavour particles (b/c/τ)
  - Need to tag them in many physics analysis
  - Unstable but flying particles

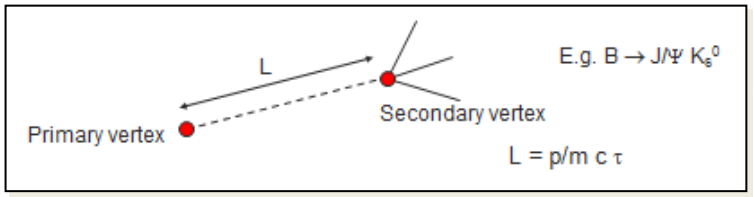
Stable particles $\tau > 10^{-6}$ s		$c\tau$
n		2.66km
μ		658m
Very long lived particles $\tau > 10^{-10}$ s		
π, $K^\pm$, K_L^0	2.6×10^{-8}	7.8m
K_S^0, $E^\pm$, Δ^0	2.6×10^{-10}	7.9cm
Long lived particles $\tau > 10^{-13}$ s		
$\tau^\pm$	0.3×10^{-12}	91μm
B_d^0, B_s^0, Δ_b	1.2×10^{-12}	350μm
Short lived particles		
π^0, η^0	8.4×10^{-17}	0.025μm
ρ,ω	4×10^{-23}	10^{-9}μm!!

+ charm (D)

$$\beta = v/c$$

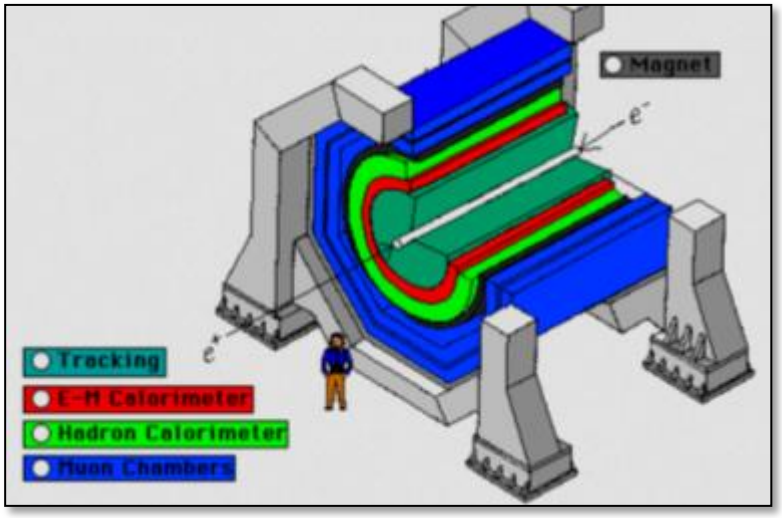
$$\langle d \rangle = \beta.\gamma.c.\tau$$

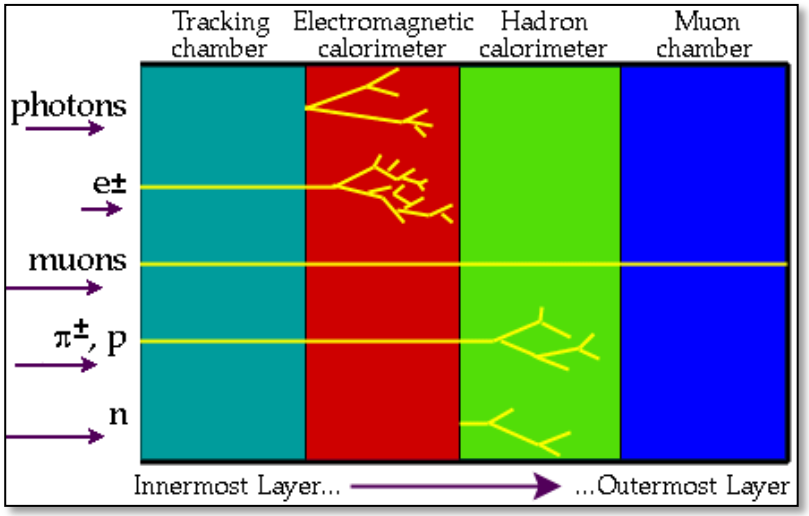
$$\gamma = \frac{1}{\sqrt{1-\beta^2}}$$



Jets containing b and c quarks
Tau leptons :
⇨ Typical <d> ~ O(10-100s μm)

⇨ Necessary resolution on these vertex position (impact parameter): ~ O(10 μm)


Magnet
Tracking
E-M Calorimeter
Hadron Calorimeter
Muon Chambers


Tracking chamber
Electromagnetic calorimeter
Hadron calorimeter
Muon chamber
photons
e±
muons
π±, p
n
Innermost Layer...
...Outermost Layer

Time resolution in the context of e^+e^- colliders

CBM-MVD
5 µs - Under development

HL-LHC
e.g.MALTA - 25 ns

STAR-PXL detector
185 µs - 2014

ALICE-ITS2
10 µs - In construction

MUPIX-X
(Mu3e)
<10 ns

ALICE-Beyond LS3

ALICE-Beyond LS4

Belle-II upgrade ?

Others (ions, pp, etc.)

100 µs | 10 µs | 1µs | 100 ns | 10 ns | 1 ns | 100 ps | 10 ps

Time resolution

ILC R&D

Minimal
VXD requirements
~1-4 µs

SIT &
Bunch tagging
300-500 ns

backscatterred filter
10 ns

Cut at 15 ns

Particle ID

ILC

VXD requirements
~1 µs

tt bunch ~ 3 µs

Z bunch
~ 20 ns

Particle ID

FCCee

CLIC 5 ns

CLIC bunch
500 ps

Particle ID

CLIC

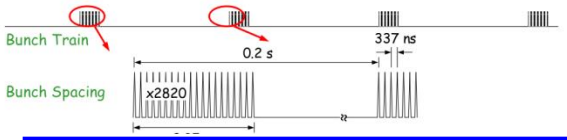
VXD requirements
~1 µs

(H) Bunch
680 ns

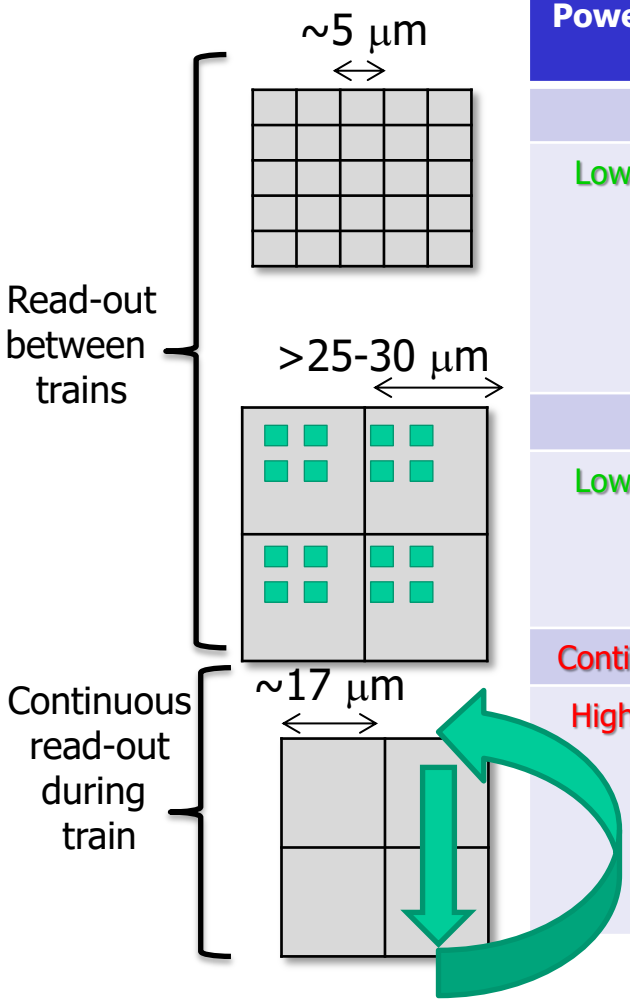
(Z) bunch
25 ns

Particle ID

CEPC



Read-out strategies and technologies



Power	Time resolution	Spatial resolution	Advantages	Caveats
Fine pixels (e.g. FPCCD)				
Low	1 complete train	~ 1 μm	Spatial Resolution Hit separation Beam background tagging capabilities ? (cluster shapes)	⇨x16 #pixels to read-out in 200ms ⇨No time stamping ⇨Occupancy issues ?
In pixel circuitry to store hits with time stamping (e.g. chronopixels, SOI)				
Low	Single or few bunches (>~ 0.5 μs)	>~ 5 μm	Hit time stamping Well suited to outer layers	⇨BX time stamping storage in conflict with granularity
Continuous read-out during train (e.g. DEPFET, CMOS): rolling shutter or priority encoding.				
High	Few to 10s bunches (1-50 μs)	~ 3 μm	Time & spatial resolution compromise	Power cycling mandatory ? ⇨F(Lorentz) ~ 10^5 grams ⇨Distribute 100s Amps shortly before train ⇨heat cycles the ladders.

Technology	FPCCD	DEPFET	SOI	CMOS	iLGAD
Added value (example)	Very granular	Low material budget	2 tier process (high density μcircuits)	Industry evolution	PID

VXD-ILD: Data flux

Layer	DBD occupancy (hits/cm²/BX)	Detector surface (mm²)	#hits/BX	#hits/read out	#hits/train	# hits/s	Data rate (Mbits/train)	Data rate (Mbits/s)	Data rate (Mbits/train) With safety factor of 3	Data rate (Mbits/s) With safety factor of 3
	@ √s = 500 GeV	Length x width x # ladders		assuming 4 μs i.e. 8 BX	Assuming 1312 bunches per train	Assuming 5 trains / s	Assuming 16 bits/pixel & 5 pixels/hit & 10 bits header = 100 bits/hit	Assuming 16 bits/pixel & 5 pixels/hit & 10 bits header = 100 bits/hit	Assuming 16 bits/pixel & 5 pixels/hit & 10 bits header = 100 bits/hit	Assuming 16 bits/pixel & 5 pixels/hit & 10 bits header = 100 bits/hit
0	6.32 ± 1.76	125 x 11 x 10 = 13 750	870	7000	1140 K	5700 K	114	570	342	1710
1	4.00 ± 1.18	125 x 11 x 10 = 13 750	550	4400	720 K	3600 K	72	360	216	1080
2	0.25 ± 0.11	125 x 2 x 22 x 11 =60 500	150	1200	197 K	985 K	19.7	98.5	59.1	295.5
3	0.21 ± 0.09	125 x 2 x 22 x 11 =60 500	130	1040	171 K	855 K	17.1	85.5	51.3	256.5
4	0.04 ± 0.03	125 x 2 x 22 x 17 =93 500	40	320	52 K	260 K	5.2	26	15.6	78
5	0.04 ± 0.03	125 x 2 x 22 x 17 =93 500	40	320	52 K	260 K	5.2	26	15.6	78
TOTAL		335 500 mm² ~ 0.35 m²	1780	14280	2332 K	11660 K	233.2	1166	700	3500

- average raw data size (without or with safety factor on beam background included)
  - Average size per BX : ~0.18 Mbits / BX ⇨ 0.54 Mbits / BX (with safety factor of 3) ~375 Gbits/s (instantaneous)
  - Average size per event (~8 BX) : ~1.4 Mbits/ readout ⇨ 4.3 Mbits / readout (with safety factor of 3)
  - Average size per train : ~233 Mbits / train ⇨ 700 Mbits / train (with safety factor of 3)
  - Average size per second : ~1166Mbits / s ⇨ 3500 Mbits / s (with safety factor of 3)

ILC beam background

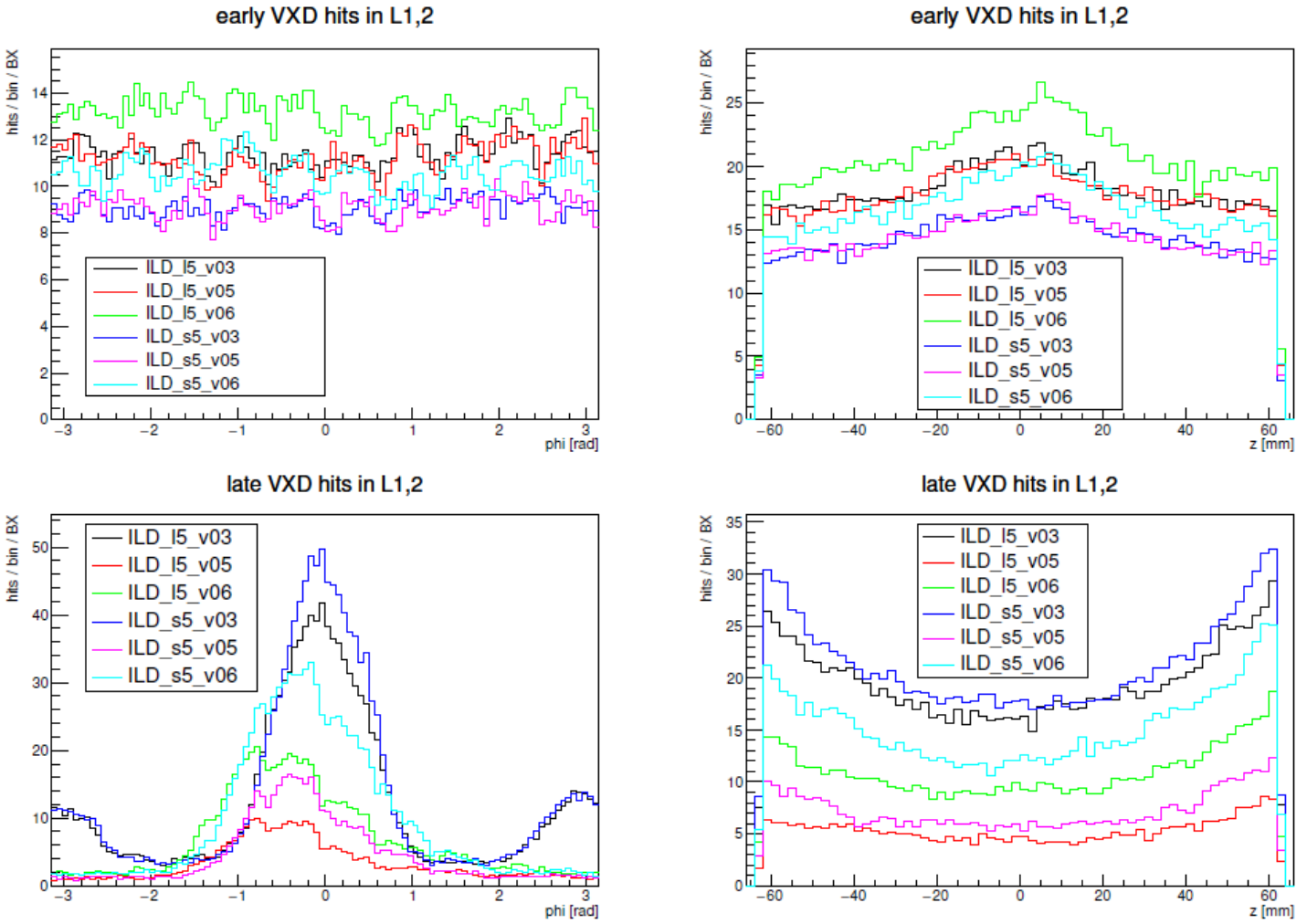
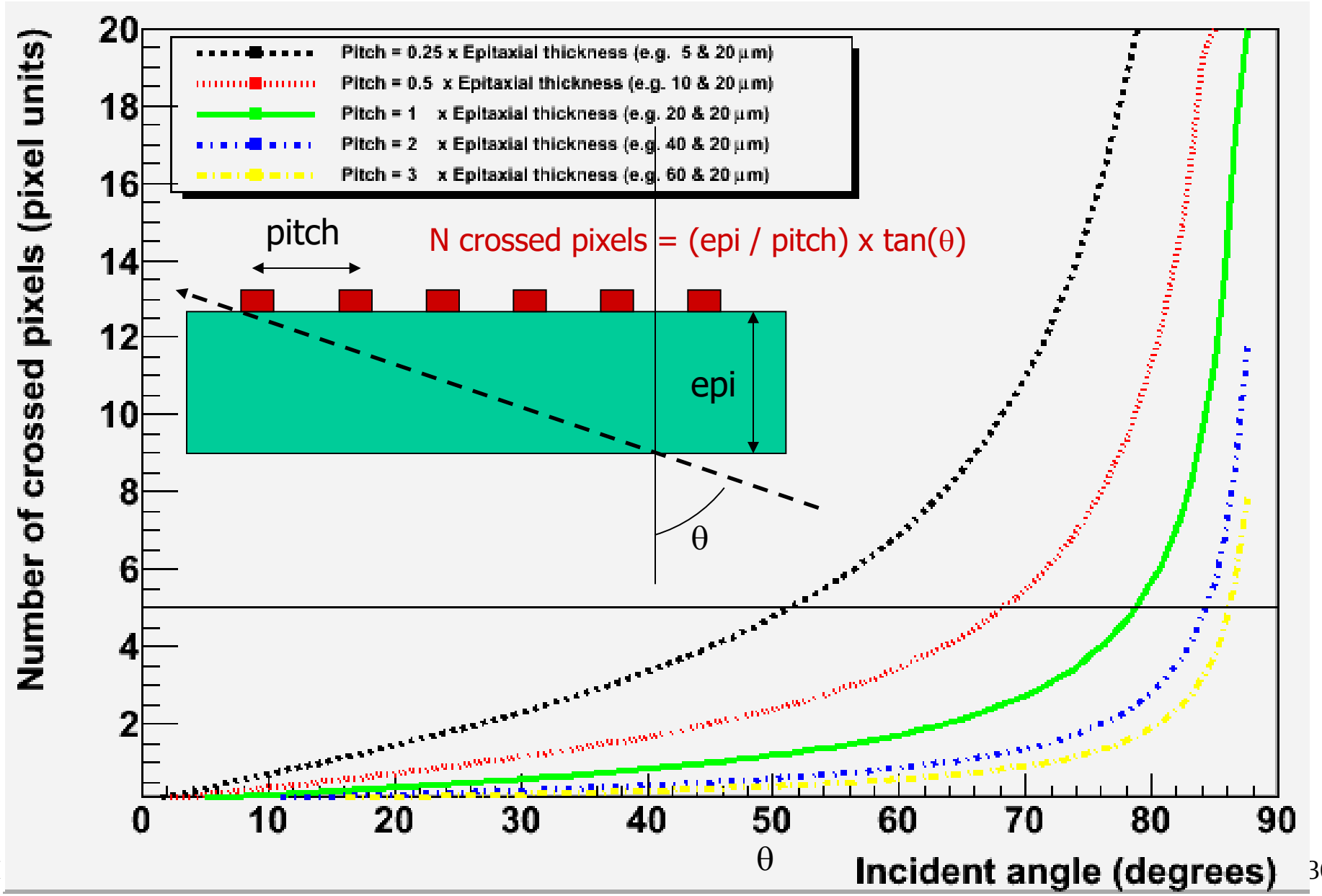


Figure 2: Distributions in azimuthal angle ϕ and z of early (upper plots) and late (lower plots) hits in the first two VXD layers.

Number of crossed pixels: pitch / epitaxial layer thickness



ALICE ITS Material budget