LABORATOIRE DE L'ACCÉLÉRATEUR LINÉAIRE

# Népal, un processeur PCIe avec 64 liens à 10 Gbps


***Pierre Matricon***

***B.Debennerot, P.Dinaucourt, JC.Hernandez, P.Rusquart, R.Sliwa, S.Trochet***



**Journées VLSI-CAO IN2P3, Marseille CPPM, 11-13 juin 2014**


cnrs IN2P3 Les deux infinis

UNIVERSITÉ PARIS SUD

# Avant-propos

**Cette présentation reflète l'expérience que j'ai eue en m'inspirant de la méthodologie proposée par Altera pour réaliser une carte mettant en œuvre un gros FPGA avec des liens gigabit.**

# Plan de l'exposé

- **Motivation, cahier des charges**
- **Choix des technologies**
  - ***FPGA Stratix 5***
  - ***Circuit imprimé spécial***
- **Étapes de conception**
  - ***Code initial du Stratix***
  - ***Schéma de la carte***
  - ***Alimentations***
  - ***Découplages***
  - ***Routage***
- **Résultats et conclusion**

# Motivation, cahier des charges

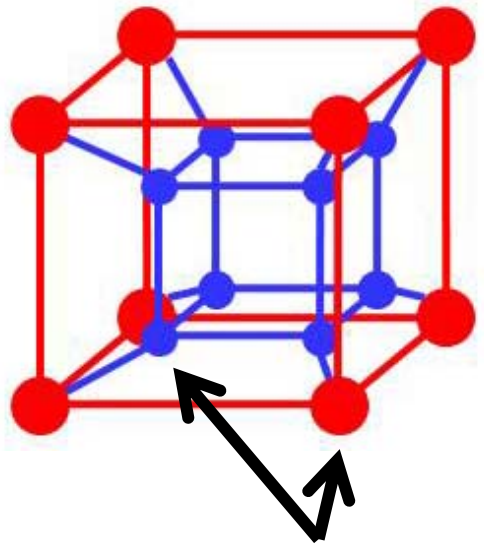
**Créer un super calculateur pour des applications impliquant des échanges incessants entre unités de calcul (puces multi cœurs, GPU ...).**
**Applications de théorie ou de modélisation telles que :**

- ***la QCD sur réseau (Lattice QCD)***
- ***le climat***
- ***l'aérodynamique***
- ***l'hydrodynamique***
- ***le feu …***

**Le calculateur est constitué de milliers, voire de millions, d'unités de calcul qui communiquent entre voisines sur une même puce, sur une carte, et entre cartes, châssis et baies (là surtout réside le problème).**

**Idée directrice : réduire le temps perdu dans les échanges grâce à des interconnexions point à point entre les cartes, châssis et baies, et grâce à un maillage serré 3D et 4D.**

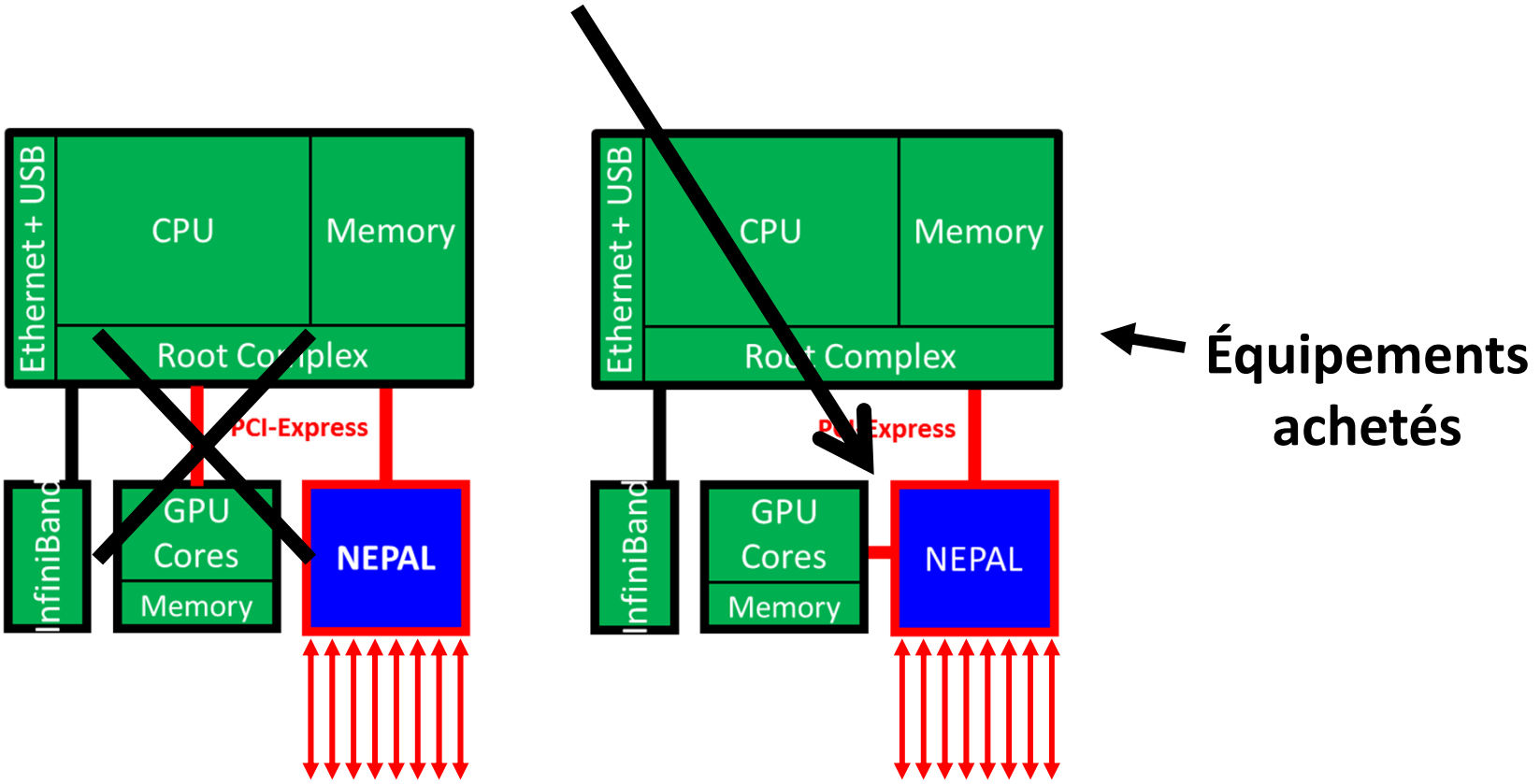
**Exemple d'une maille élémentaire en 4D**



**unités de calcul**

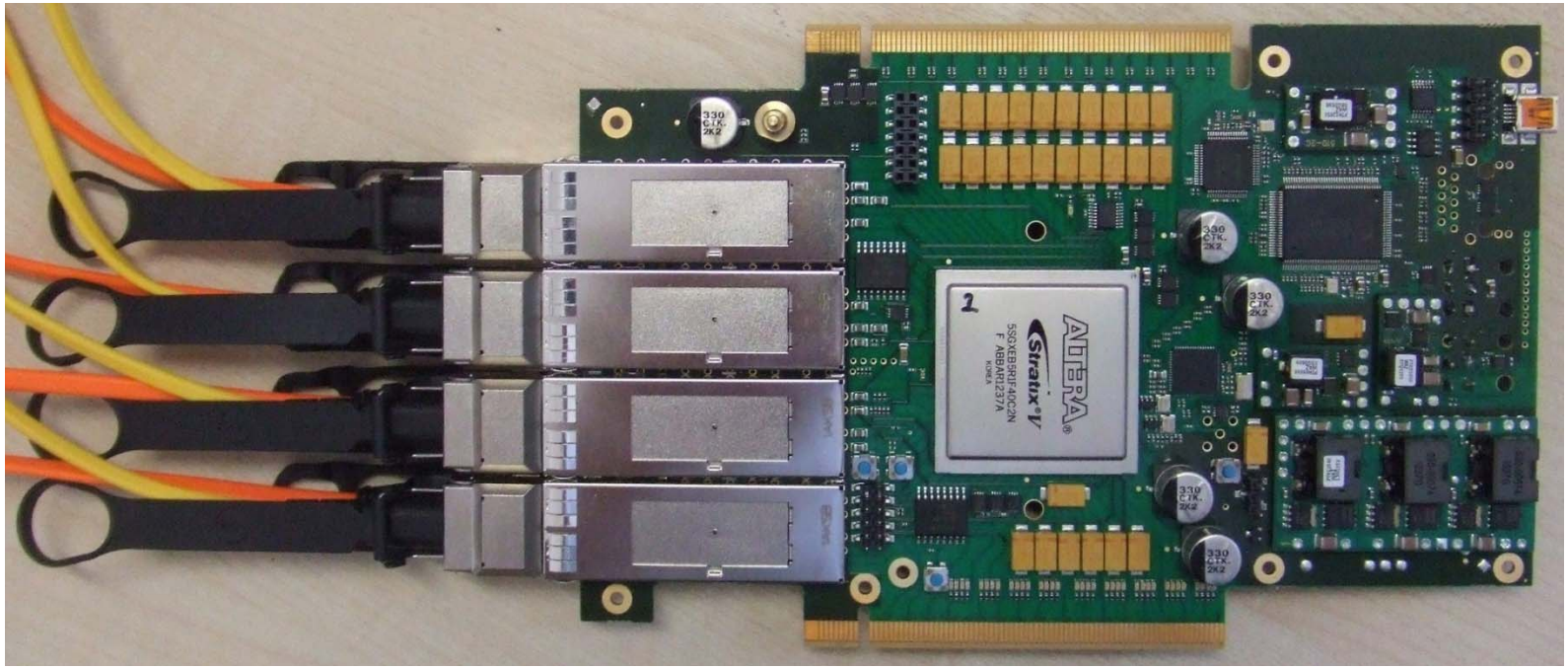
**Dans notre projet, l'unité de calcul est basée sur des cartes de calcul PCIe et des équipements commerciaux.**

**La carte Népal sert à relier physiquement les cartes de calcul PCIe, sans passer par le CPU.**



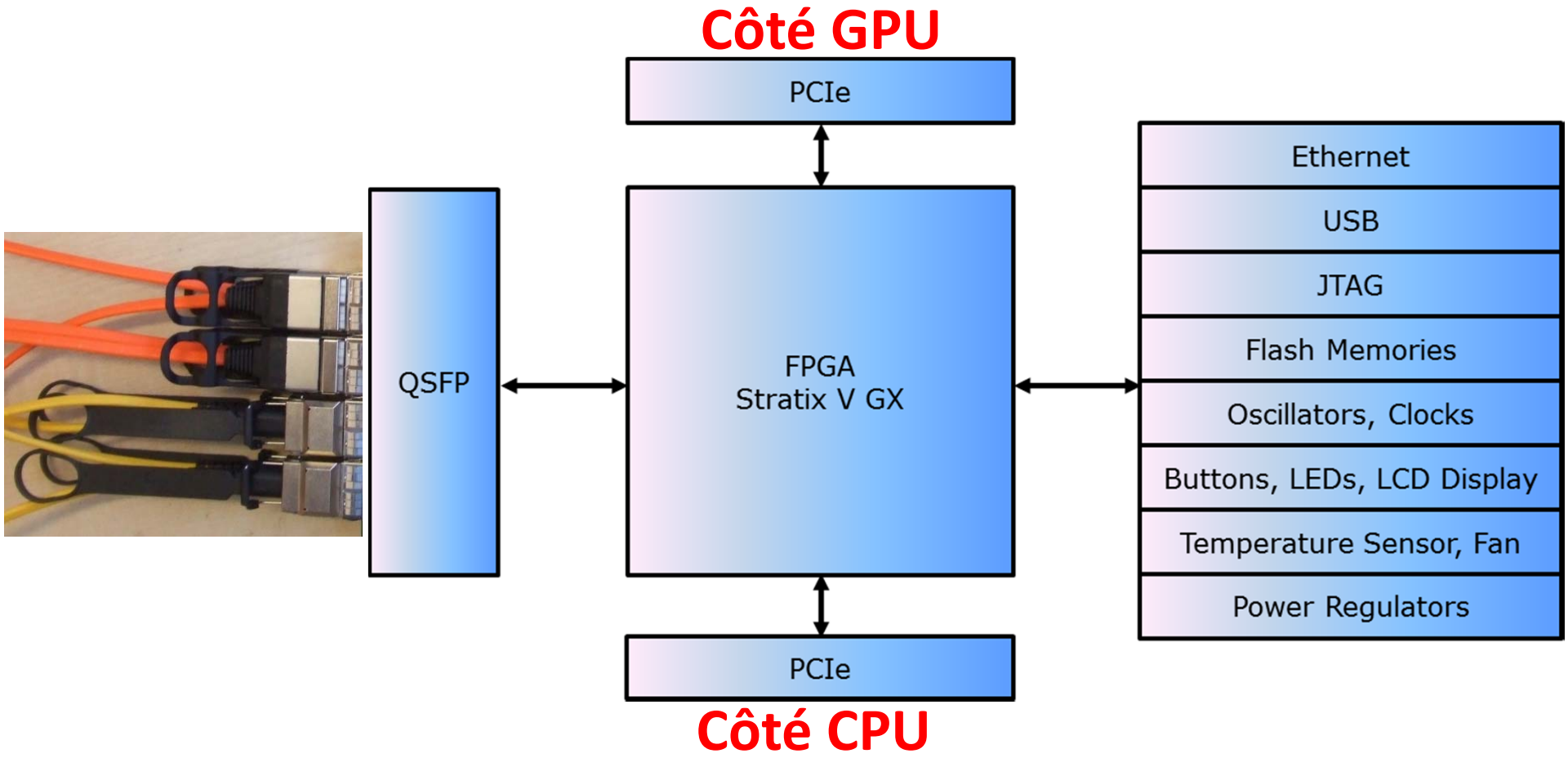
# Choix des technologies
## FPGA Stratix 5

**Nous avons choisi le Stratix 5SGXEB5R1 qui dispose de 66 liens offrant chacun jusqu'à 14 G bits par seconde**

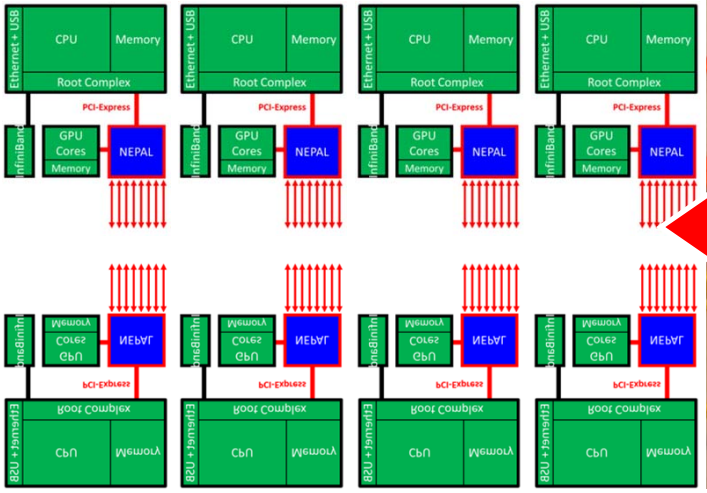


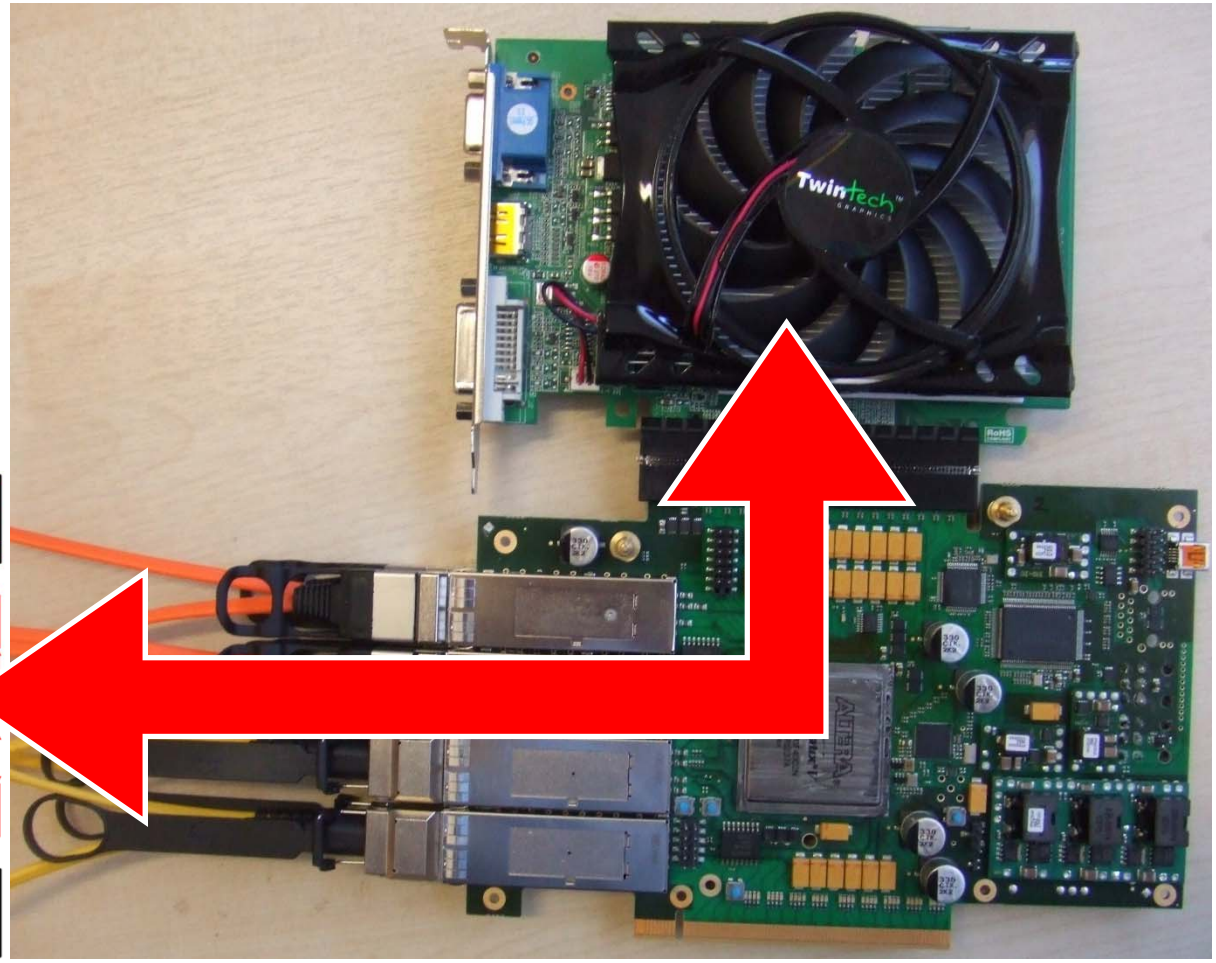

**8 câbles optiques actifs (64 fibres)**
**relient une carte à chacune des 8 voisines**

# Architecture de la carte Népal

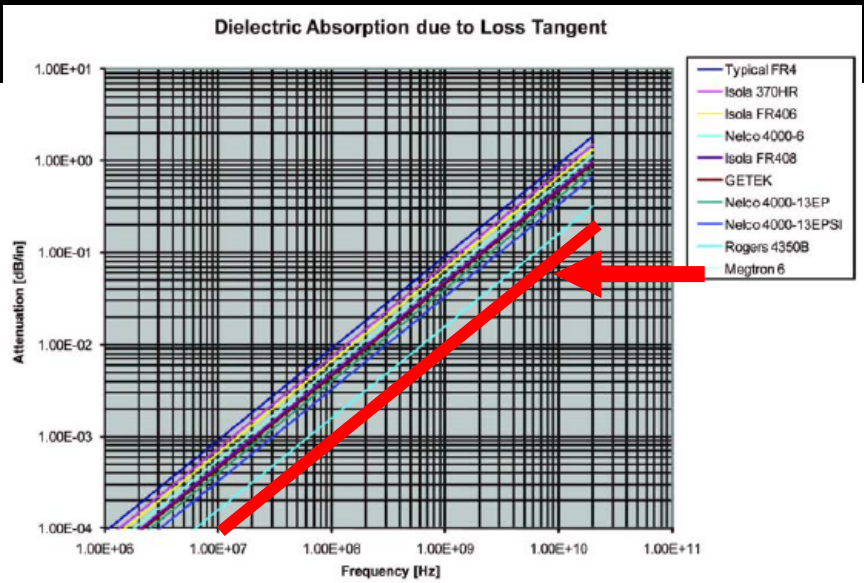


# La carte de calcul est montée directement sur la carte Népal et communique avec chaque voisine en point à point.



# Choix des technologies
## Circuit imprimé spécial




**Après le FPGA, la matière du circuit imprimé est choisie pour les hautes fréquences. En effet, un signal qui se propage à haute vitesse dans une ligne de transmission perd de la puissance et de l'amplitude.**

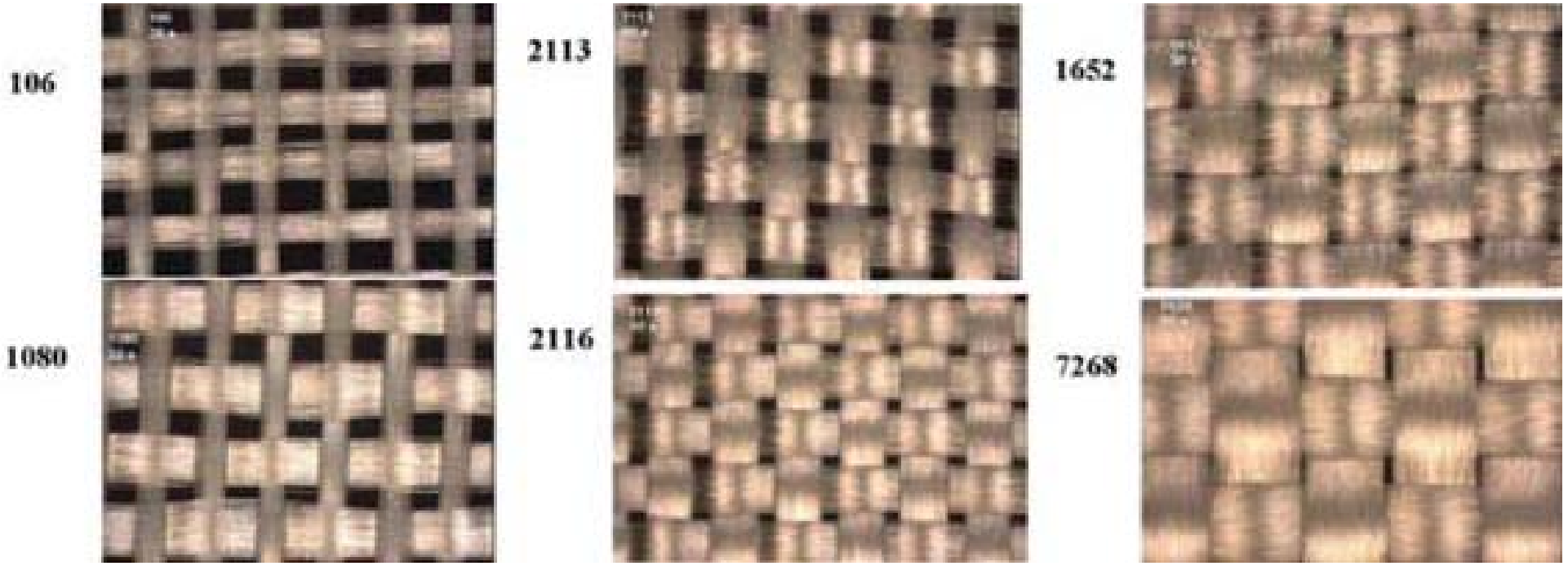
**Le paramètre le plus important pour les hautes fréquences est le facteur de dissipation Df appelé aussi tangente de perte.**

**Nous avons choisi le Megtron 6.**

| Material | εr | Df |
|---|---|---|
| Typical FR4 | 4 | 0.02 |
| GETEK | 3.9 | 0.01 |
| Isola 370HR | 4.17 | 0.016 |
| Isola FR406 | 4.29 | 0.014 |
| Isola FR408 | 3.70 | |
| Megtron 6 | 3.4 | 0.002 |
| Nelco 4000-6 | 4.12 | |
| Nelco 4000-13 EP | 3.7 | 0.009 |
| Nelco 4000-13 EP SI | 3.2 | 0.008 |
| Rogers 4350B | 3.48 | 0.0037 |

**Pour les lignes de transmission à hautes fréquences, outre Df, il est aussi utile de prendre en compte :**

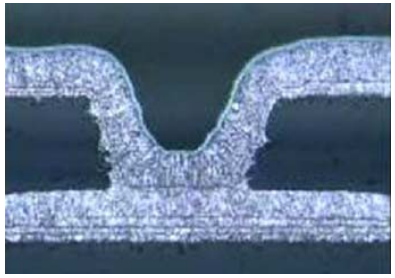
- les pertes résistives, donc la résistance de la ligne avec l'effet de peau
- le εr du matériau (plus bas le εr, plus haute l'impédance)
- le tissage des fibres du matériau pour l'homogénéité

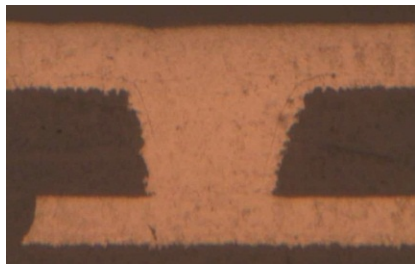


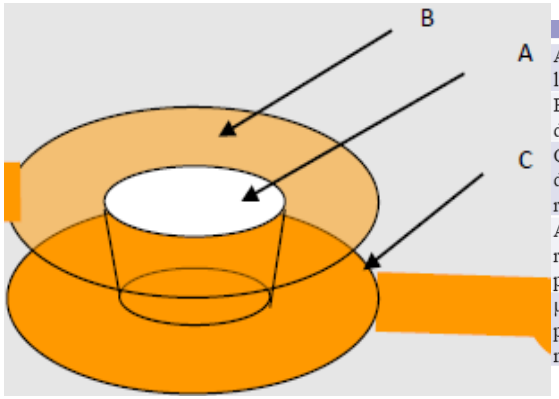
# Choix des technologies
## Vias laser dans les pads de brasure

**Nous utilisons les vias-in-pad laser essentiellement pour accéder aux billes du BGA.**

**Vias-in-pad signifie que les trous micro vias sont positionnés directement dans les pads de brasure. Le dépôt de brasage remplit le creux.**



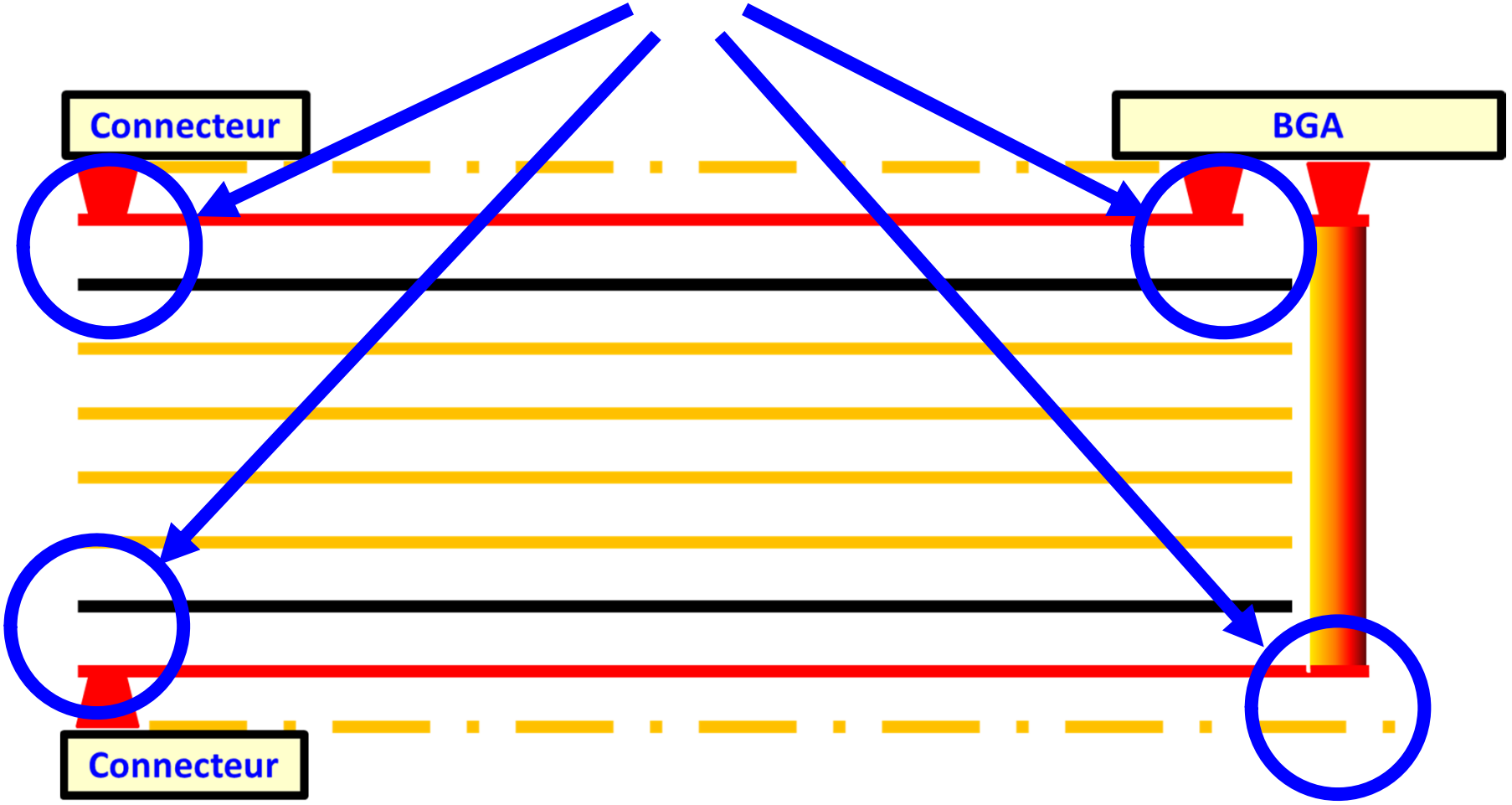





| | µvias laser | | | | |
|---|---|---|---|---|---|
| A Perçage laser | 100 µm | 130 µm | 150 µm | 220 µm | > 220 possible |
| B Pastilles d' émission | 250 µm | 300 µm | 320 µm | 400 µm | Trou + 170 µm |
| C Pastilles de réception | 250 µm | 300 µm | 320 µm | 400 µm | Trou + 170 µm |
| Aspect ratio preferred | 0,85 | 0,85 | 0,85 | 0,85 | 0,85 |
| µvia filling profondeur maxi | 85 µm | 110 µm | 125 µm | 180 µm | |

**La caractéristique la plus importante pour les micro vias est l'aspect-ratio, c.à.d. le rapport diamètre sur profondeur.**

**Les vias laser sont notamment utiles pour les lignes rapides qui ne doivent pas présenter de bouts parasites (stubs).**



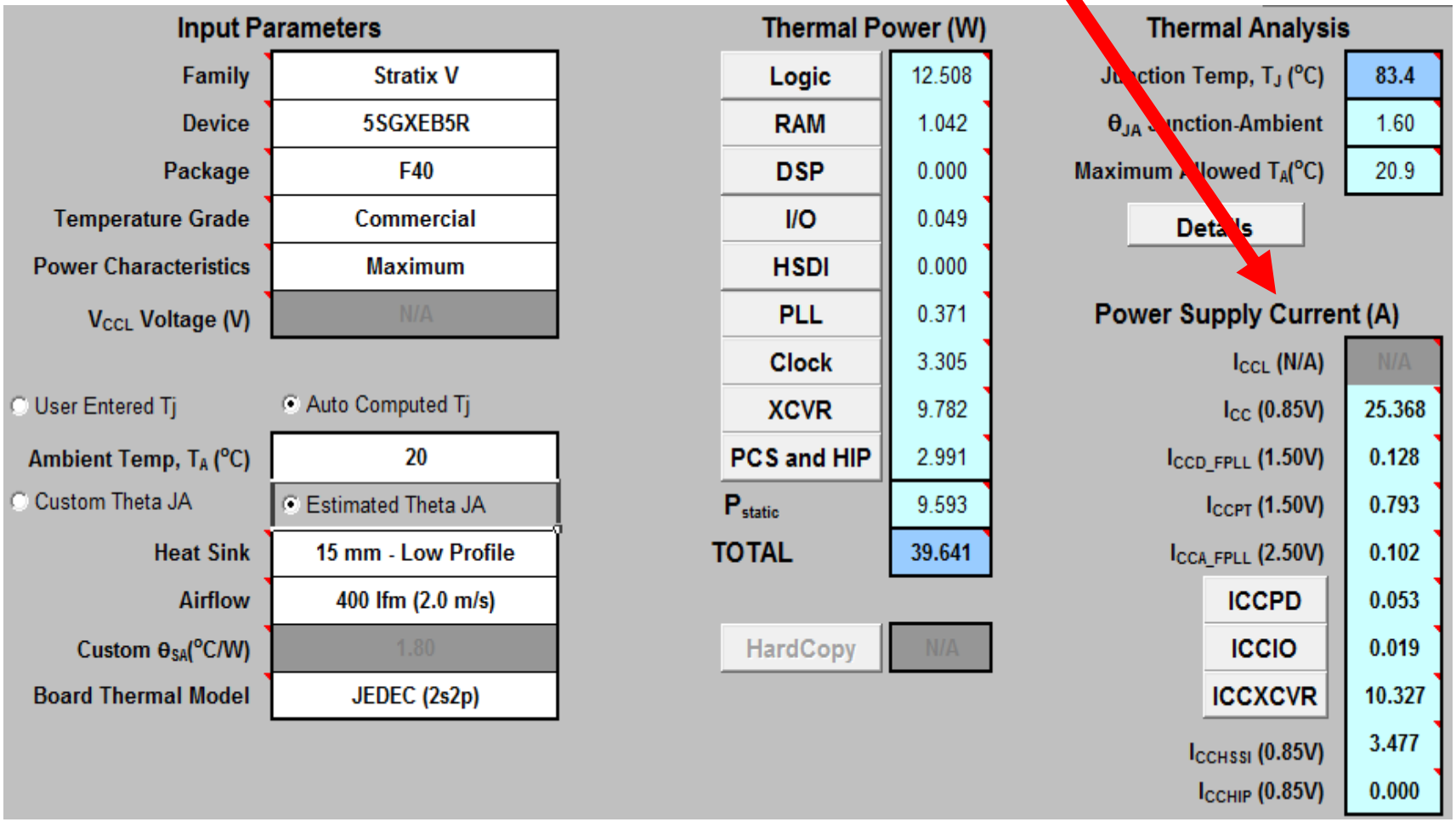
# Étapes de conception

- Code initial du Stratix
- Schéma de la carte
- Alimentations

- **Le code initial du Stratix permet de valider l'emplacement des entrées et sorties, notamment des ports rapides.**

- **Avec le schéma de la carte, le code initial permet aussi d'évaluer les ressources à mobiliser dans le Stratix pour réaliser les fonctions nécessaires au projet (PLLs, horloges, mémoires, logique, etc.).**

**L'évaluation des ressources sert à compléter la feuille de calcul Early Power Estimator qui fournit les éléments permettant de construire le réseau de distribution de puissance (courant pour chaque tension).**

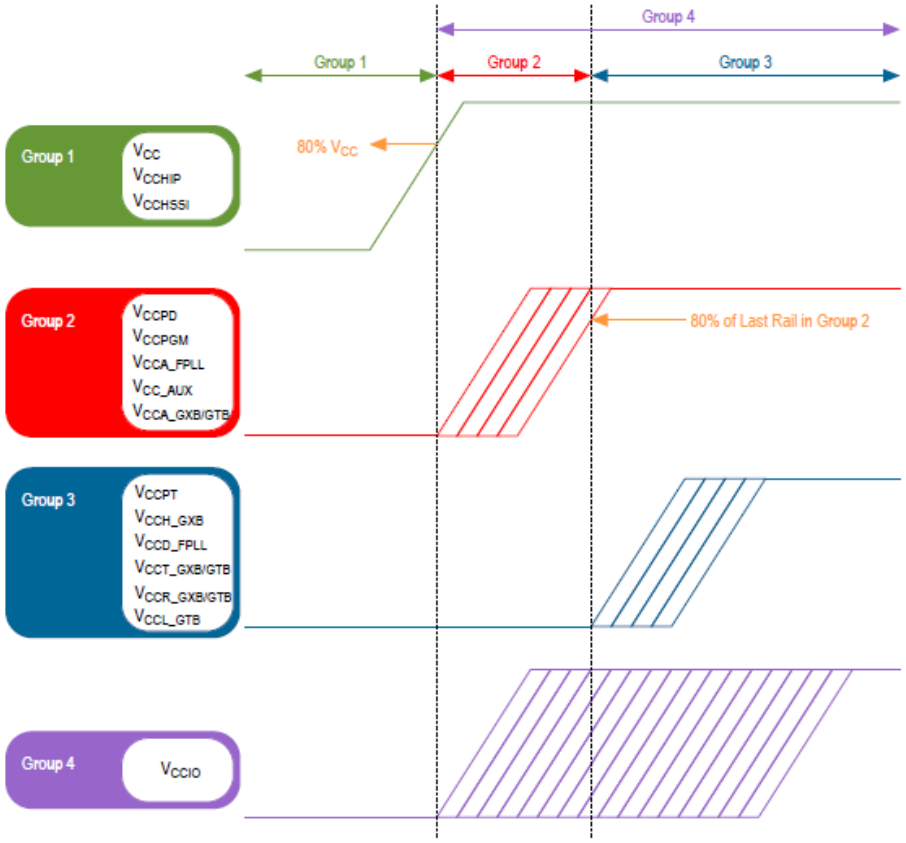


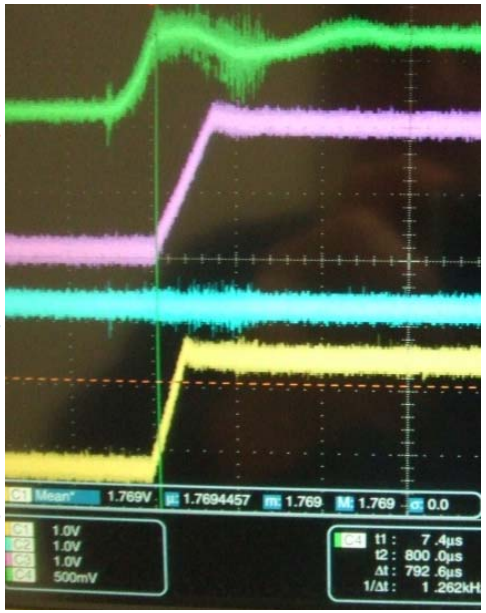
# Les régulateurs sont choisis en fonction :

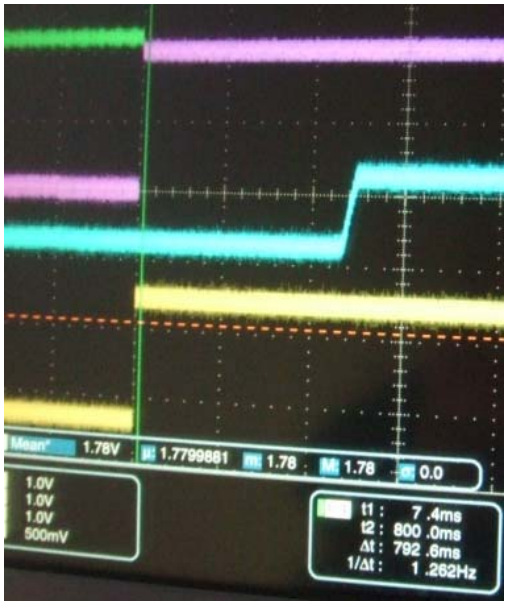
- du courant prévu pour les 15 tensions d'alimentation,
- de la précision requise,
- et du taux d'ondulation résiduelle

| Rail Name | Default voltage (V) | Allowable Ripple | Transient Current Percentage (%) | Description |
|---|---|---|---|---|
| VCC | 0.85 | 5 | 50 | Core |
| VCCHIP | 0.85 | 5 | 50 | PCIE Hard IP (Digital) |
| VCCHSSI | 0.85 | 5 | 50 | PCS Power |
| VCCIO | 1.2 - 3.0 | 5 | 100 | I/O Bank |
| VCCPD | 2.5 / 3.0 | 5 | 50 | I/O pre-drivers |
| VCCPGM | 1.8 / 2.5 / 3.0 | 5 | 50 | Programming Power |
| VCC_AUX | 2.5 / 3.0 | 5 | 50 | Programmable Power Tech Aux |
| VCCBAT | 1.2 | 5 | 100 | Battery Back-up Power Supply |
| VCCA_FPLL | 2.5 | 5 | 10 | PLL (Analog) |
| VCCD_FPLL | 1.5 | 5 | 10 | PLL (Digital) |
| VCCR_GXB | 0.85 / 1.0 | 3 | 30 | Transceiver RX (Analog) |
| VCCT_GXB | 0.85 / 1.0 | 2 | 50 | Transceiver TX (Analog) |
| VCCA_GXB | 3 | 5 | 10 | Transceiver/CDB (Analog) |
| VCCH_GXB | 1.5 | 3 | 15 | Transceiver I/O Buffer Block |
| VCCPT | 1.5 | 5 | 33 | Programmable Power Tech |

# Les 15 alimentations du Stratix sont séquencées.



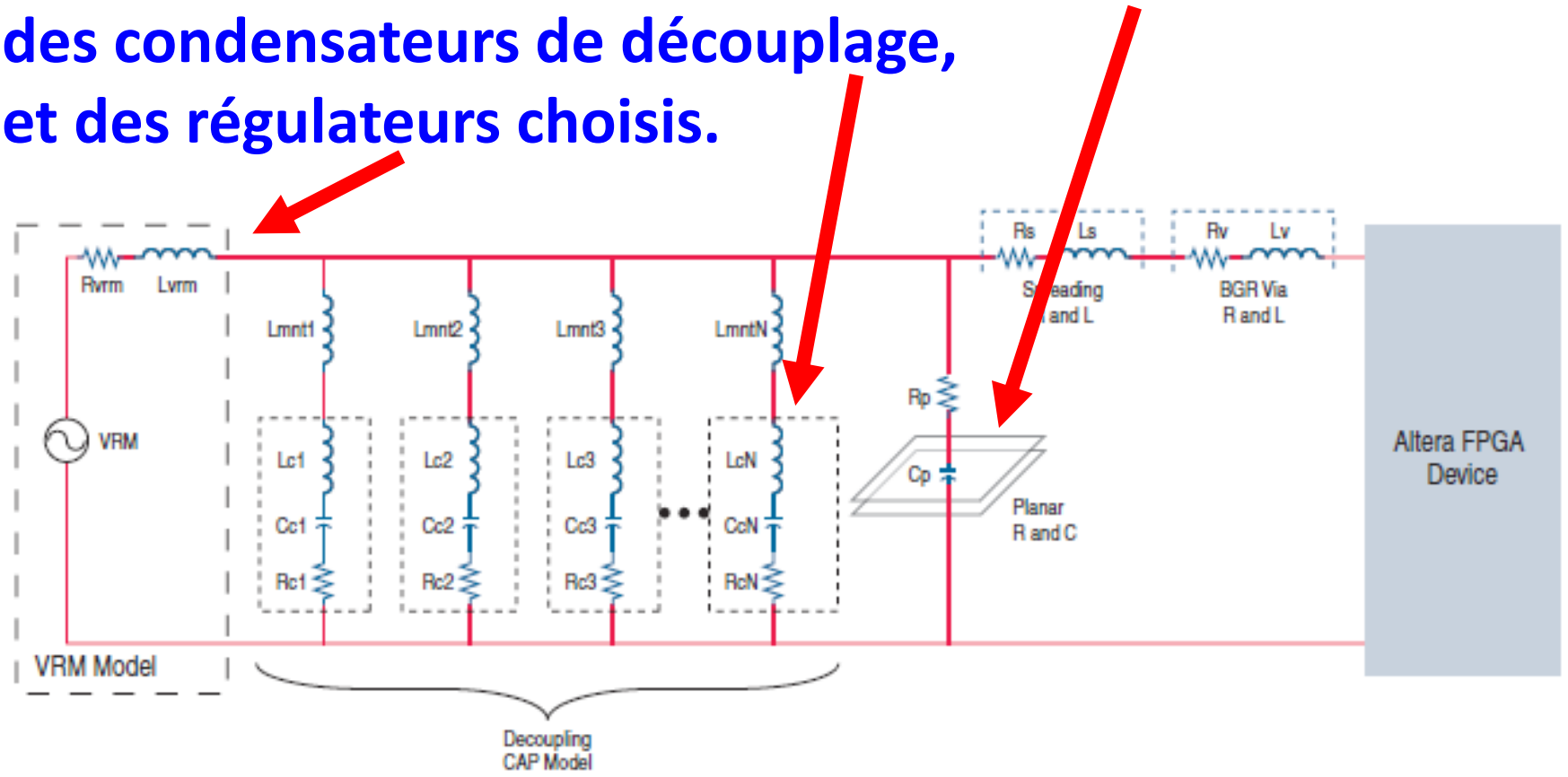





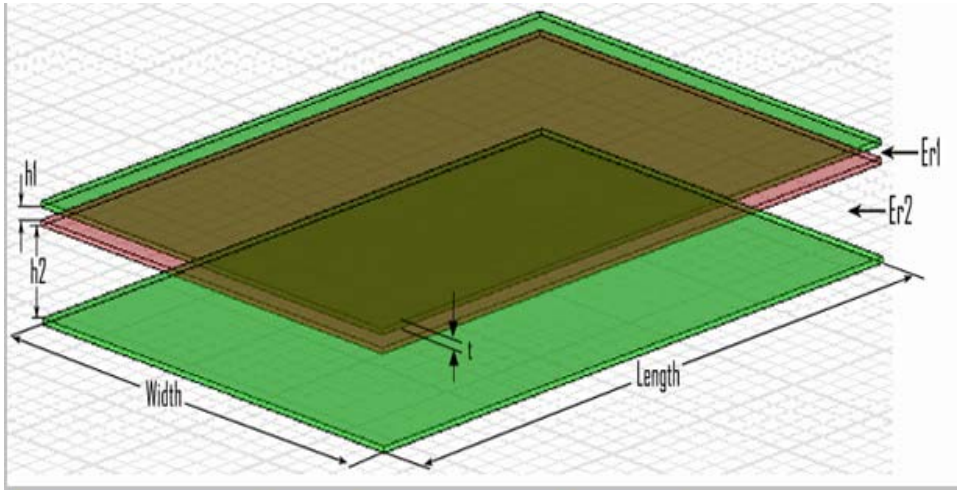
# Étapes de conception
## Découplages

**La qualité de la distribution de puissance dans la plage des fréquences utiles dépend**

- **de la capacité entre les plans d'alimentation,**
- **des condensateurs de découplage,**
- **et des régulateurs choisis.**

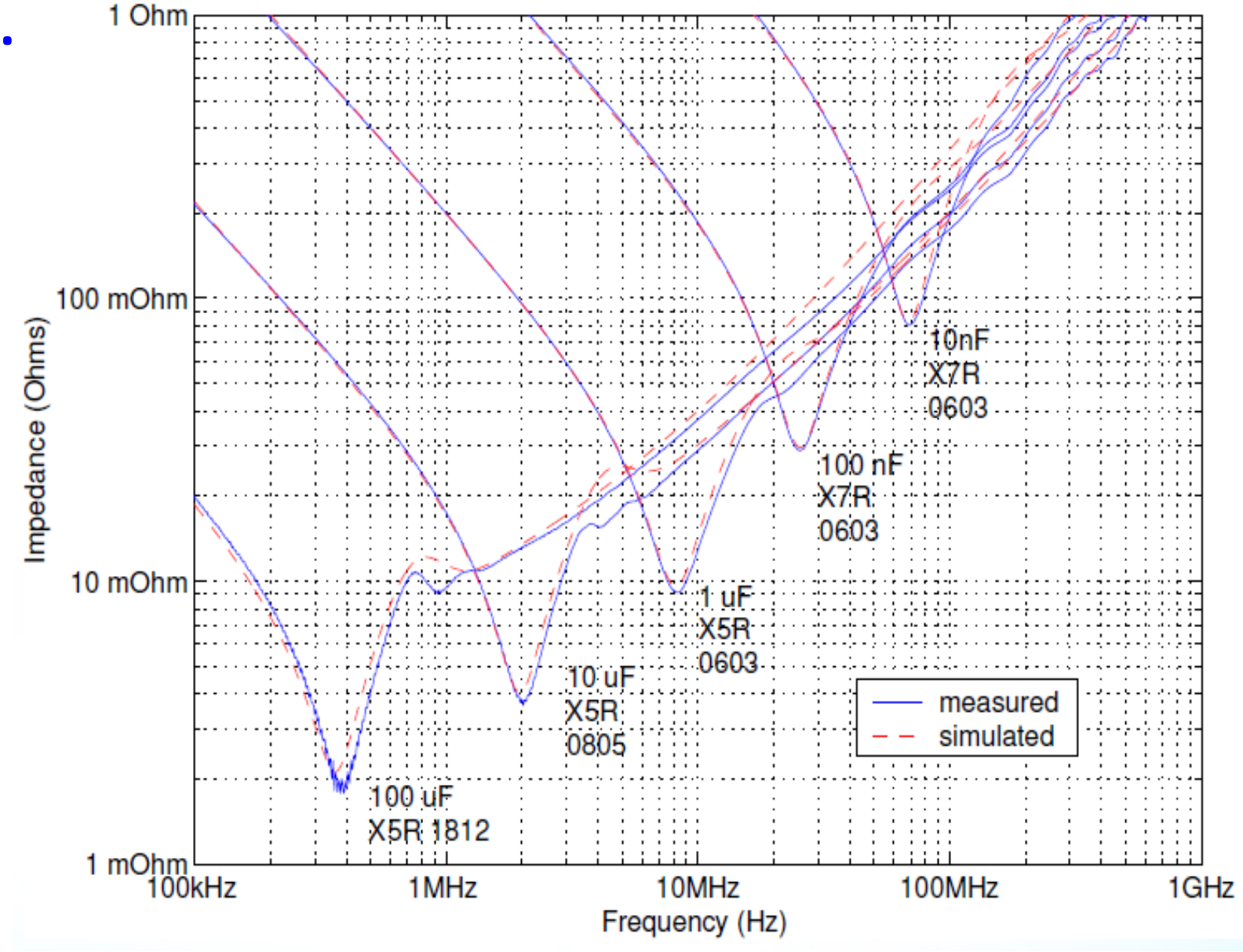


**Par exemple, sur la carte Nepal, le condensateur constitué des plans d'alimentation GND et VCC=0,9V a une capacité de 4,1 nF .**




| Planar Capacitance | Symbol | Unit | Value |
|---|---|---|---|
| Plane length | Length | mils | 7800 |
| Plane width | Width | mils | 3900 |
| Metal thickness | t | mils | 0.546 |
| Height to 1st GND plane | h1 | mils | 7.800 |
| Height to 2nd GND plane | h2 | mils | 22.152 |
| Dielectric material 1 | Er1 | Rogers 4350B | 3.48 |
| Dielectric material 2 | Er2 | Rogers 4350B | 3.48 |
| Plane capacitance 1 | C1 | µF | 0.0031 |
| Plane capacitance 2 | C2 | µF | 0.0011 |
| Total planar capacitance | Ctotal | µF | 0.0041 |
| Total sheet resistance | Rtotal | Ω | 0.0013 |

**L'impédance des condensateurs dépend de leur nature (matériau, constitution, etc.), de leur valeur et de la fréquence.**



- **Quels condensateurs ?**
- **De quelles valeurs ?**
- **Combien ?**

**Pour calculer la nature des condensateurs de découplage, leurs valeurs et leurs nombres, une méthode adéquate est fournie par le tableur PDN d'Altéra qui utilise la méthode de l'impédance cible dans le domaine fréquentiel.**

$$Z_{TARGET} = \frac{VoltageRail\ (\%Ripple/100)}{MaxTransientCurrent}$$

**Par exemple, pour la tension de cœur VCC=0,9V du Stratix avec 5% de ripple,**
**pour un courant de 30 A avec 50% de courant transitoire, l'impédance cible est de 3 mΩ.**

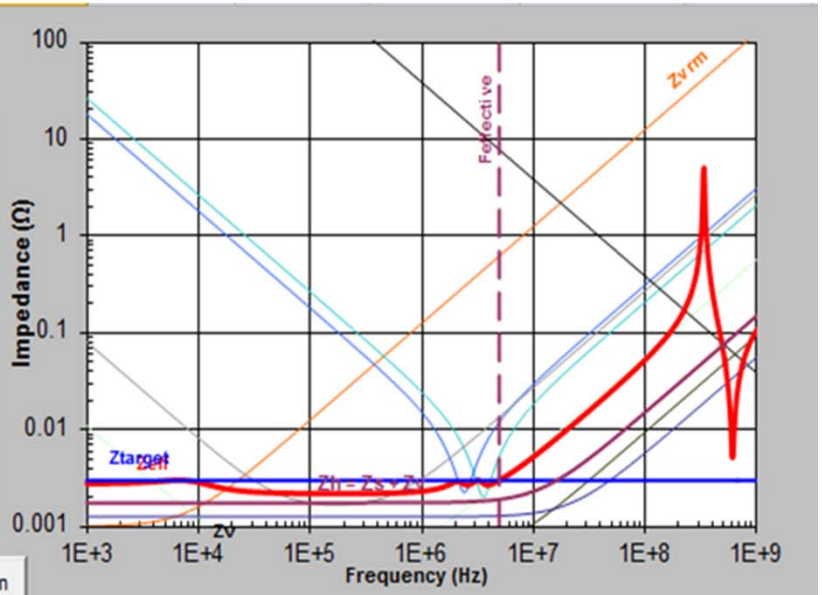
$$Z_{TARGET} = \frac{VoltageRail\ (\%Ripple/100)}{MaxTransientCurrent}$$

# Le tableur conseille d'utiliser les 63 condensateurs indiqués.

| | |
|---|---|
| Family / Device | Stratix_V |
| Available Devices | 5SGXEB5R_F40 |
| Power Supply Rail | VCC |

| Summary | Options | R (Ω) | L (nH) | C (µF) |
|---|---|---|---|---|
| VRM | Switcher | 1.0E-03 | 2.0E+01 | N/A |
| Spreading | Low | 0.0005 | 0.0150 | N/A |
| BGA Via | Calculate | 0.0001 | 0.0089 | N/A |
| Plane Cap | Calculate | 0.0013 | 0.023 | 0.0041 |

| Target Impedance | | Units | Value | Legend |
|---|---|---|---|---|
| Supply Voltage (Min) | | V | 0.9 | N/A |
| I max | | A | 30 | N/A |
| Transient Current | | % | 50 | N/A |
| Vripple (+/-) | | % | 5 | N/A |
| Feffective | Calculate | MHz | 4.9 | Feffective |
| Ztarget = ΔV / ΔI | | Ω | 0.0030 | Ztarget |




Auto Decouple | Verify Solution

**Decoupling Caps**

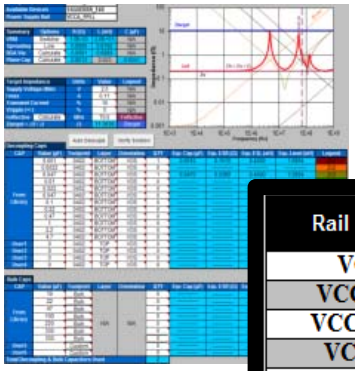
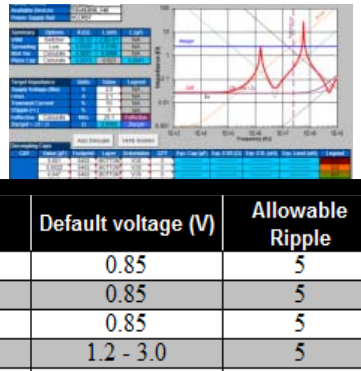
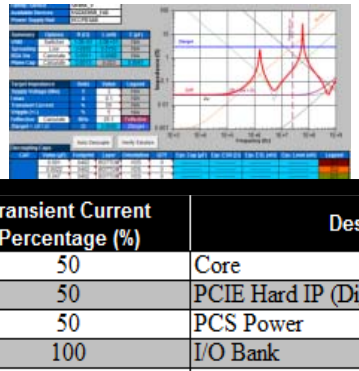
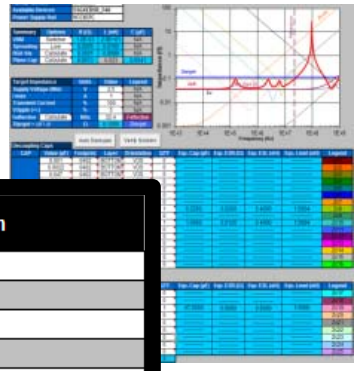
| CAP | Value (µF) | Footprint | Layer | Orientation | QTY | Eqv. Cap (µF) | Eqv. ESR (Ω) | Eqv. ESL (nH) | Eqv. Lmnt (nH) | Legend |
|---|---|---|---|---|---|---|---|---|---|---|
| From Library | 0.001 | 0402 | BOTTOM | VOS | 0 | ---------- | ---------- | ---------- | ---------- | Zc1 |
| | 0.0022 | 0402 | BOTTOM | VOS | 0 | ---------- | ---------- | ---------- | ---------- | Zc2 |
| | 0.047 | 0402 | BOTTOM | VOS | 0 | ---------- | ---------- | ---------- | ---------- | Zc3 |
| | 0.01 | 0402 | BOTTOM | VOS | 0 | ---------- | ---------- | ---------- | ---------- | Zc4 |
| | 0.022 | 0402 | BOTTOM | VOS | 0 | ---------- | ---------- | ---------- | ---------- | Zc5 |
| | 0.047 | 0402 | BOTTOM | VOS | 0 | ---------- | ---------- | ---------- | ---------- | Zc6 |
| | 0.1 | 0402 | BOTTOM | VOS | 0 | ---------- | ---------- | ---------- | ---------- | Zc7 |
| | 0.22 | 0402 | BOTTOM | VOS | 0 | ---------- | ---------- | ---------- | ---------- | Zc8 |
| | 0.47 | 0402 | BOTTOM | VOS | 0 | ---------- | ---------- | ---------- | ---------- | Zc9 |
| | 1 | 0402 | BOTTOM | VOS | 6 | 6.0000 | 0.0020 | 0.0667 | 0.2649 | Zc10 |
| | 2.2 | 0402 | BOTTOM | VOS | 4 | 8.8000 | 0.0023 | 0.1000 | 0.3974 | Zc11 |
| | 4.7 | 0402 | BOTTOM | VOS | 0 | ---------- | ---------- | ---------- | ---------- | Zc12 |

**Bulk Caps**

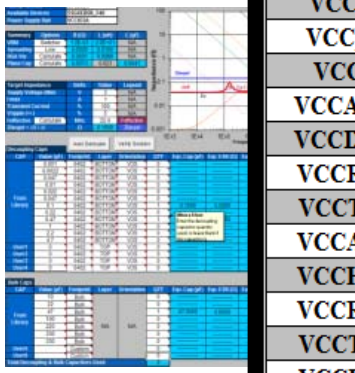
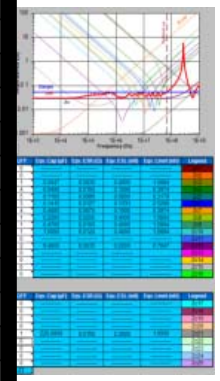
| CAP | Value (µF) | Footprint | Layer | Orientation | QTY | Eqv. Cap (µF) | Eqv. ESR (Ω) | Eqv. ESL (nH) | Eqv. Lmnt (nH) | Legend |
|---|---|---|---|---|---|---|---|---|---|---|
| From Library | 10 | Bulk | N/A | N/A | 0 | ---------- | ---------- | ---------- | ---------- | Zc17 |
| | 22 | Bulk | | | 0 | ---------- | ---------- | ---------- | ---------- | Zc18 |
| | 47 | Bulk | | | 0 | ---------- | ---------- | ---------- | ---------- | Zc19 |
| | 100 | Bulk | | | 0 | ---------- | ---------- | ---------- | ---------- | Zc20 |
| | 220 | Bulk | | | 9 | 1980.0000 | 0.0017 | 0.2556 | 0.1778 | Zc21 |
| | 330 | Bulk | | | 44 | 14520.0000 | 0.0006 | 0.0523 | 0.0386 | Zc22 |
| | 330 | Bulk | | | 0 | ---------- | ---------- | ---------- | ---------- | Zc23 |

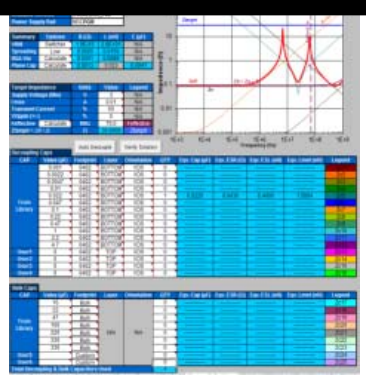
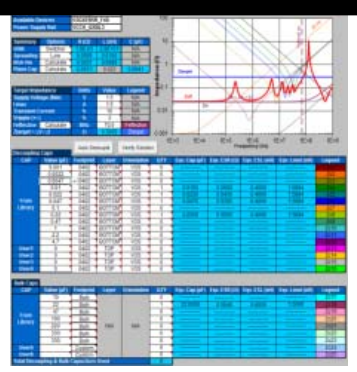
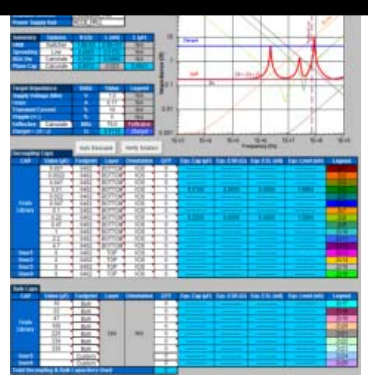
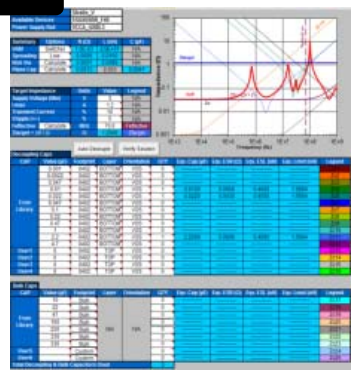
| Total Decoupling & Bulk Capacitors Used | 63 |
|---|---|

# Pour chaque tension, le tableur fournit la liste des condensateurs recommandés.

| Rail Name | Default voltage (V) | Allowable Ripple | Transient Current Percentage (%) | Description |
|---|---|---|---|---|
| VCC | 0.85 | 5 | 50 | Core |
| VCCHIP | 0.85 | 5 | 50 | PCIE Hard IP (Digital) |
| VCCHSSI | 0.85 | 5 | 50 | PCS Power |
| VCCIO | 1.2 - 3.0 | 5 | 100 | I/O Bank |
| VCCPD | 2.5 / 3.0 | 5 | 50 | I/O pre-drivers |
| VCCPGM | 1.8 / 2.5 / 3.0 | 5 | 50 | Programming Power |
| VCC_AUX | 2.5 / 3.0 | 5 | 50 | Programmable Power Tech Aux |
| VCCBAT | 1.2 | 5 | 100 | Battery Back-up Power Supply |
| VCCA_FPLL | 2.5 | 5 | 10 | PLL (Analog) |
| VCCD_FPLL | 1.5 | 5 | 10 | PLL (Digital) |
| VCCR_GXB | 0.85 / 1.0 | 3 | 30 | Transceiver RX (Analog) |
| VCCT_GXB | 0.85 / 1.0 | 2 | 50 | Transceiver TX (Analog) |
| VCCA_GXB | 3 | 5 | 10 | Transceiver/CDB (Analog) |
| VCCH_GXB | 1.5 | 3 | 15 | Transceiver I/O Buffer Block |
| VCCR_GTB | 1 | 3 | 30 | 28G Transceiver RX (Analog) * |
| VCCT_GTB | 1 | 2 | 50 | 28G Transceiver TX (Analog) * |
| VCCL_GTB | 1 | 2 | 30 | 28G Transceiver Clock (Analog) * |
| VCCPT | 1.5 | 5 | 33 | Programmable Power Tech |

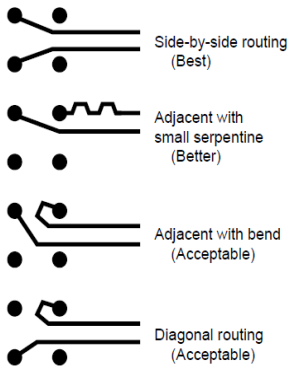



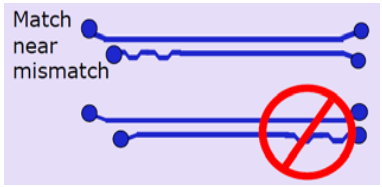
# Étapes de conception
## Routage

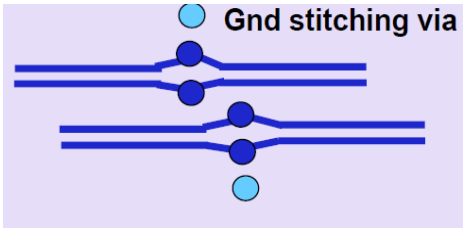
**Routage adapté aux liaisons rapides :**

- plans d'alimentation et GND proches
- alimentations sensibles sous le FPGA près du top
- condensateurs de découplage près du FPGA
- minimiser la capacité des vias => petits pads
- supprimer tous les pads non fonctionnels
- minimiser les stubs des vias
- ajouter des vias GND près des vias
- ajuster les longueurs des paires différentielles





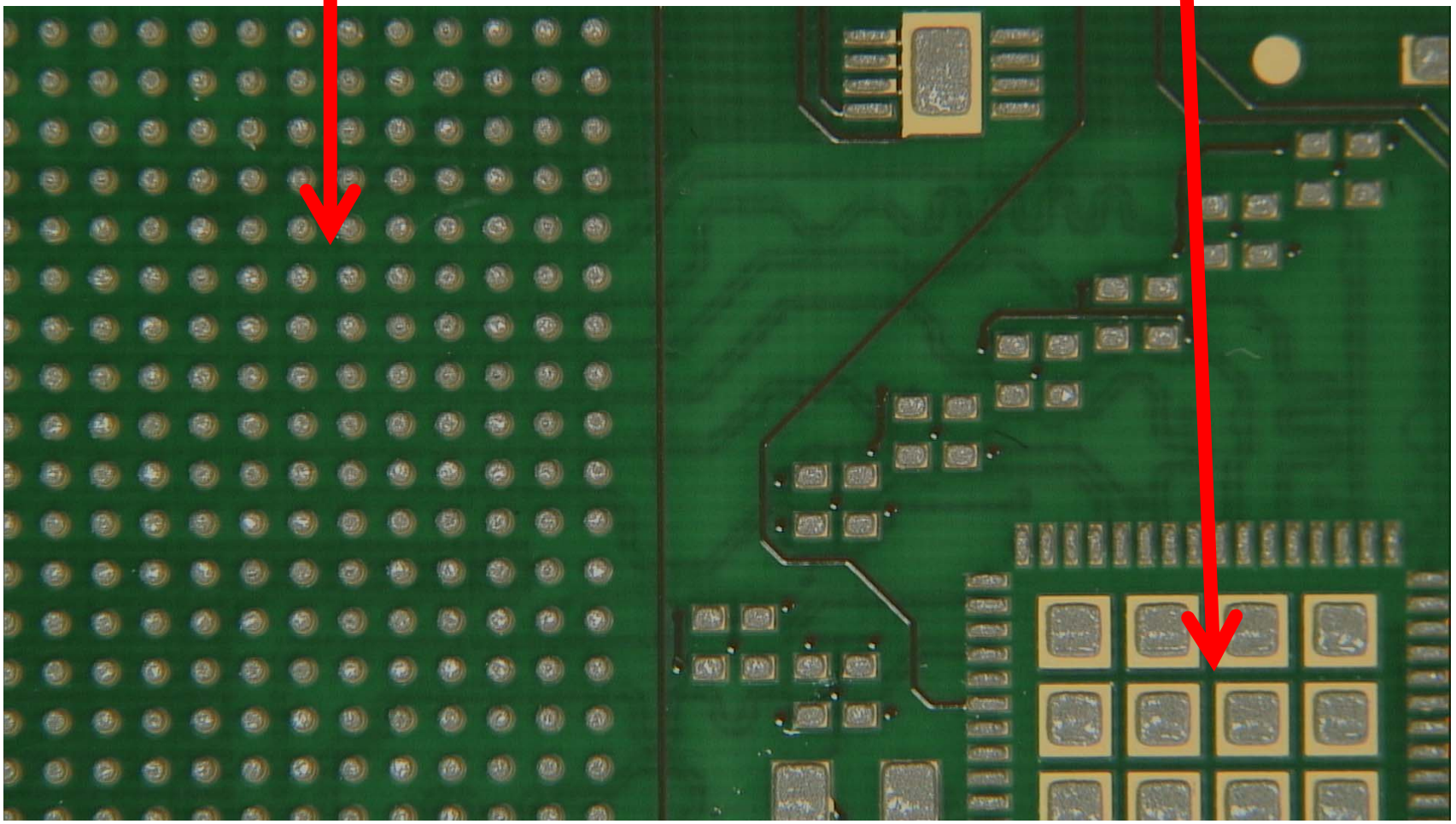


**Sur la carte Népal, les vias sont presque tous des vias-in-pad.**

Emplacement du BGA

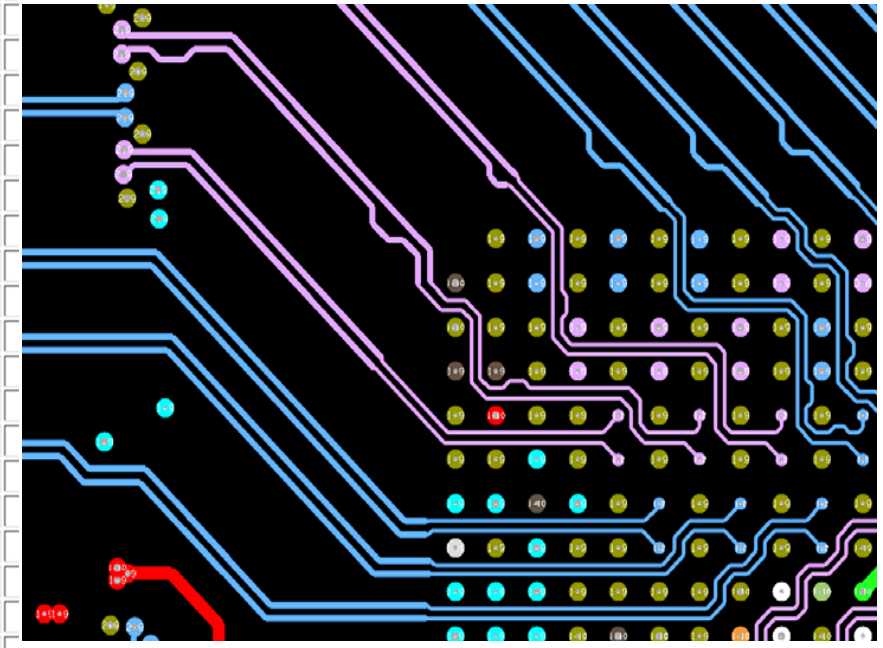
Emplacement
du générateur d'horloges
à très faible jitter

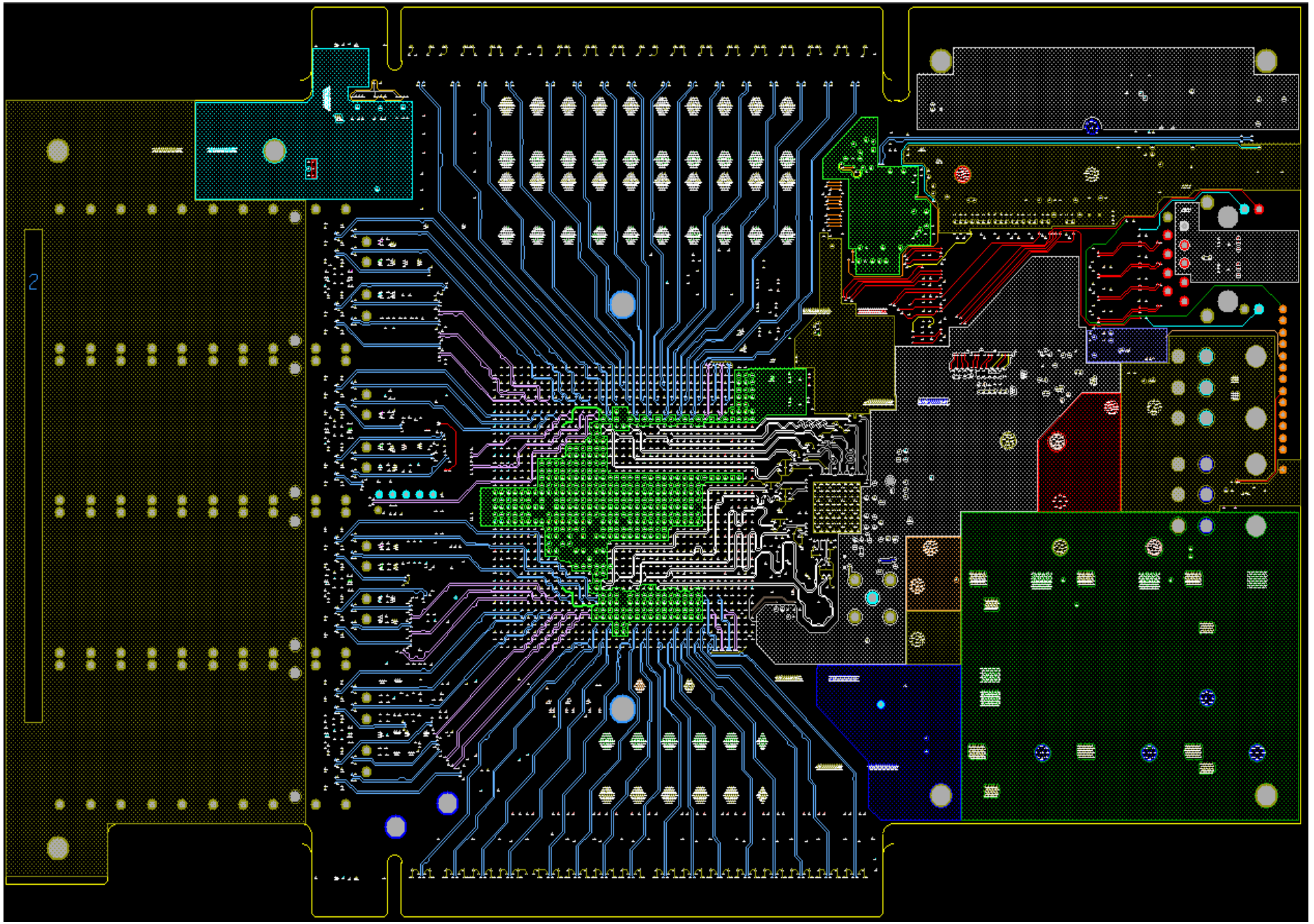


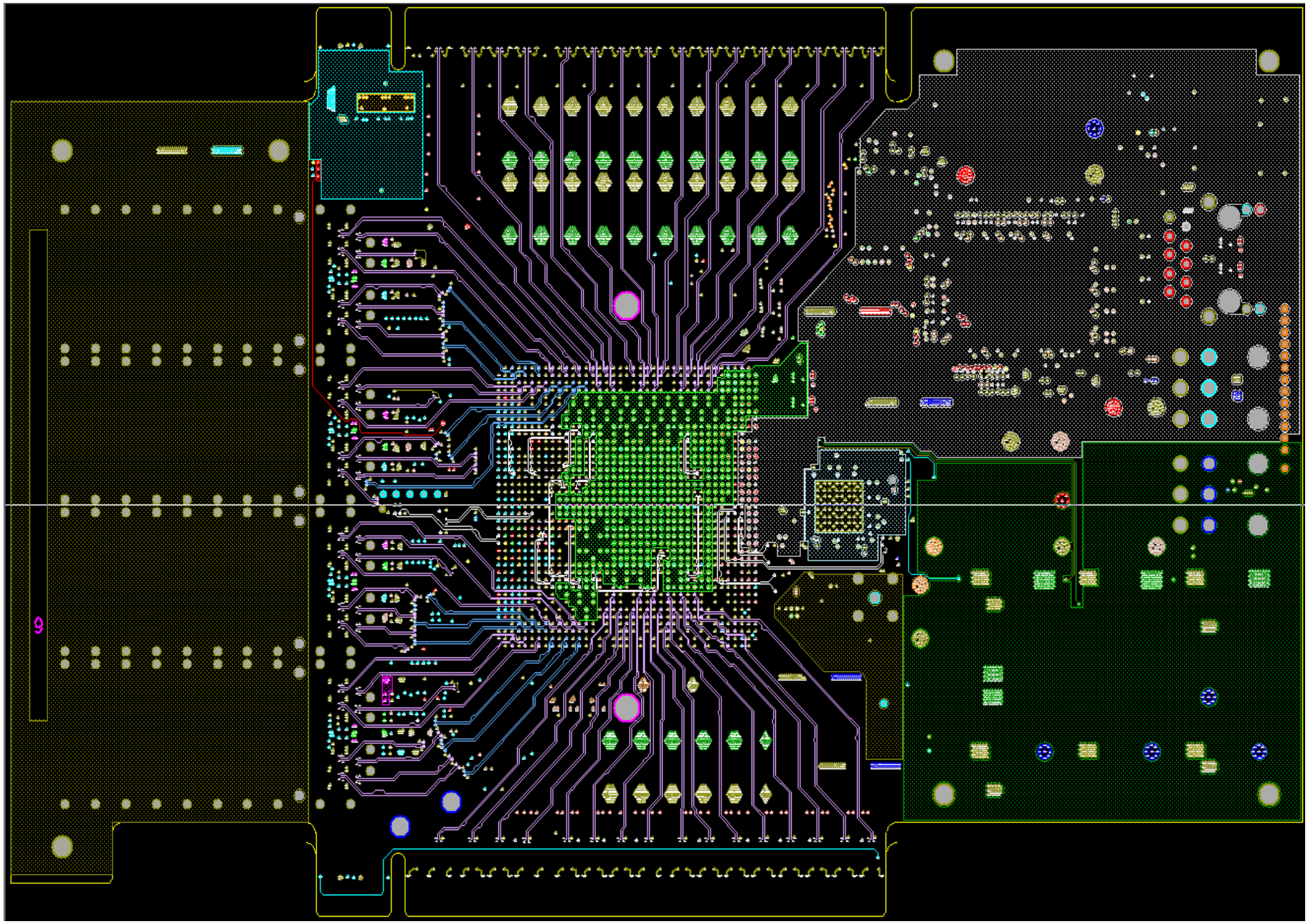
# Les longueurs de pistes ont été ajustées tout du long à 30 µm près en tenant compte de la direction du signal.

| | |
|---|---|
| QSFPR01N | 49.227 |
| QSFPR01P | 49.236 |
| QSFPR02N | 49.093 |
| QSFPR02P | 49.068 |
| QSFPR03N | 49.107 |
| QSFPR03P | 49.099 |
| QSFPR04N | 53.468 |
| QSFPR04P | 53.687 |
| QSFPR11N | 49.271 |
| QSFPR11P | 49.282 |
| QSFPR12N | 48.865 |
| QSFPR12P | 48.893 |
| QSFPR13N | 49.397 |
| QSFPR13P | 49.395 |
| QSFPR14N | 53.737 |
| QSFPR14P | 53.721 |
| QSFPR21N | 53.758 |
| QSFPR21P | 53.794 |
| QSFPR22N | 53.722 |
| QSFPR22P | 53.700 |
| QSFPR23N | 53.596 |
| QSFPR23P | 53.631 |
| QSFPR24N | 53.779 |
| QSFPR24P | 53.779 |
| QSFPR31N | 53.593 |
| QSFPR31P | 53.551 |
| QSFPR32N | 53.631 |
| QSFPR32P | 53.599 |
| QSFPR33N | 53.699 |
| QSFPR33P | 53.698 |
| QSFPR34N | 53.739 |
| QSFPR34P | 53.709 |

| | |
|---|---|
| QSFPR41N | 47.718 |
| QSFPR41P | 47.680 |
| QSFPR42N | 47.658 |
| QSFPR42P | 47.657 |
| QSFPR43N | 47.690 |
| QSFPR43P | 47.726 |
| QSFPR44N | 51.605 |
| QSFPR44P | 51.604 |
| QSFPR51N | 47.819 |
| QSFPR51P | 47.816 |
| QSFPR52N | 47.826 |
| QSFPR52P | 47.796 |
| QSFPR53N | 47.571 |
| QSFPR53P | 47.564 |
| QSFPR54N | 51.593 |
| QSFPR54P | 51.579 |
| QSFPR61N | 48.639 |
| QSFPR61P | 48.682 |
| QSFPR62N | 48.608 |
| QSFPR62P | 48.606 |
| QSFPR63N | 48.567 |
| QSFPR63P | 48.522 |
| QSFPR64N | 51.656 |
| QSFPR64P | 51.607 |
| QSFPR71N | 48.682 |
| QSFPR71P | 48.680 |
| QSFPR72N | 48.636 |
| QSFPR72P | 48.633 |
| QSFPR73N | 48.297 |
| QSFPR73P | 48.343 |
| QSFPR74N | 51.662 |
| QSFPR74P | 51.663 |






9

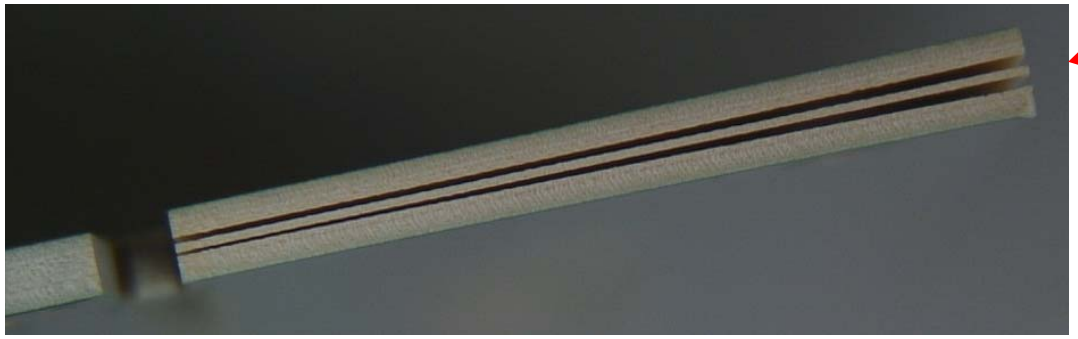
# Résultats et conclusion

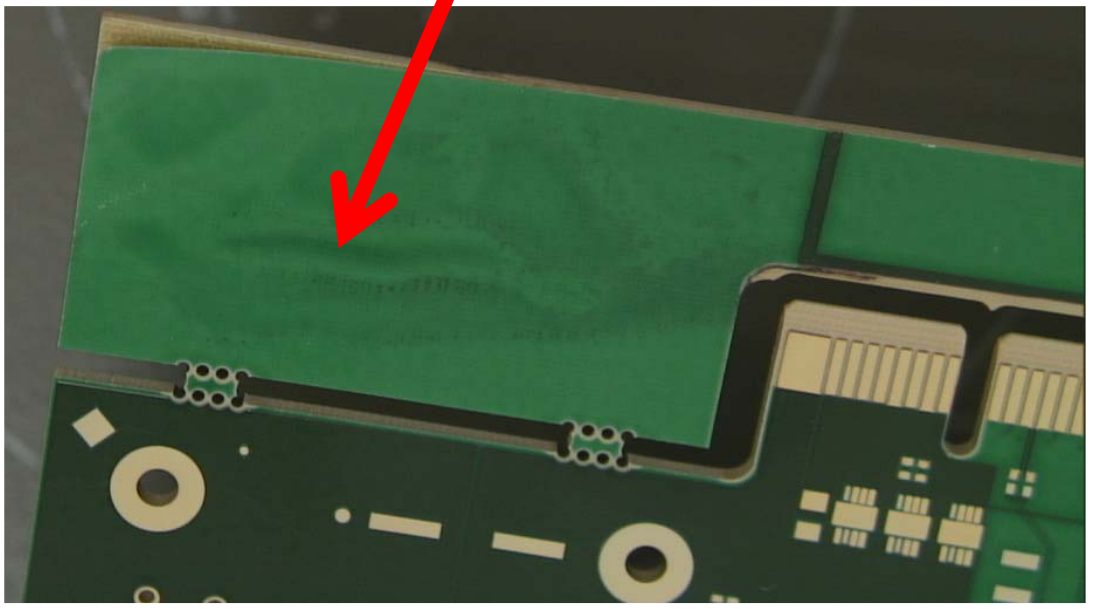
**Un projet avec quelques difficultés techniques spécifiques**

- **Composant Stratix avec une grande densité de contacts (BGA de 1517 billes au pas de 1 mm)**
- **Nombre élevé de canaux à haut débit (128 canaux à 14 giga bits/s)**
- **Exigences sur le circuit imprimé**
- **Séquencement de 15 tensions à fort courant**

# Un projet avec quelques difficultés de fabrication

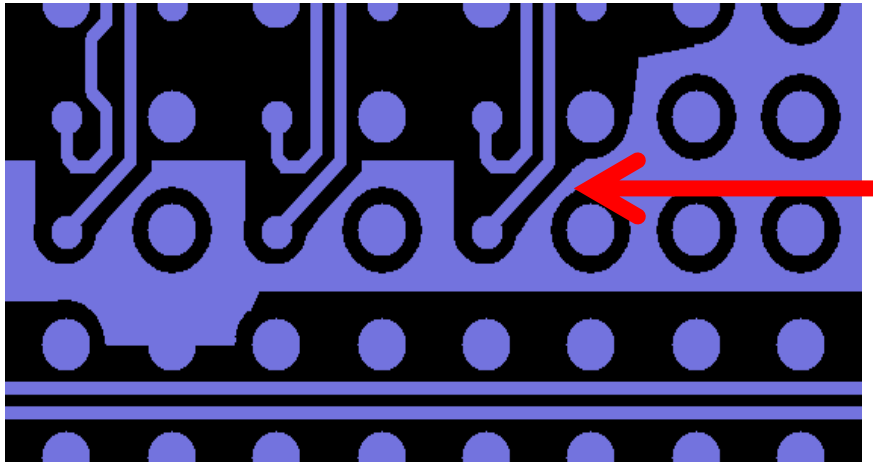
## Pour le circuit imprimé, des problèmes de délaminage :



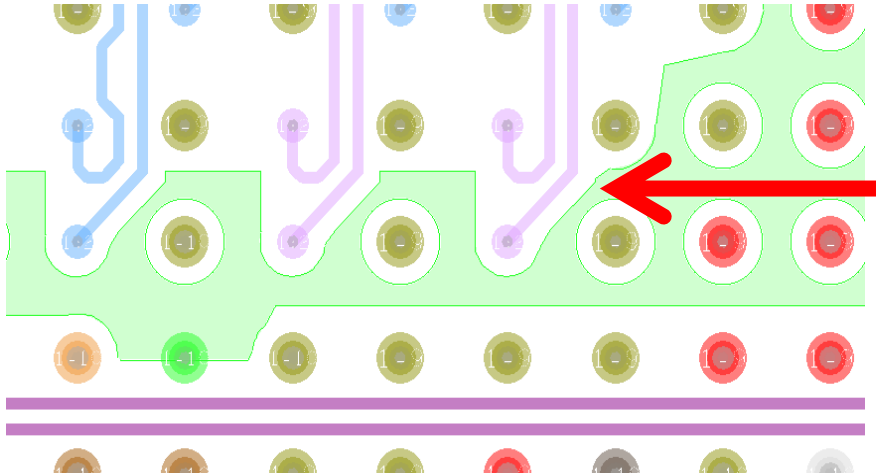


## Et des "anormalités" :

Gerber fabricant

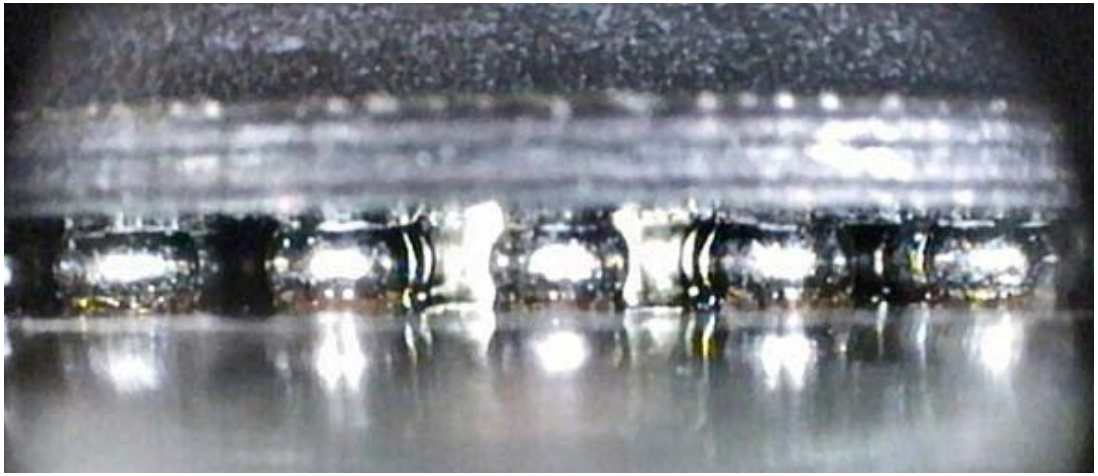


Gerber LAL



Cadence LAL

**Pour le brasage des Stratix, il a fallu s'y prendre à plusieurs reprises pour les souder.**



# Conclusion : des premiers résultats encourageants