Global is the New Local:

FPGA Architecture at 5nm and Beyond

S. Nikolić, F. Catthoor*, Z. Tőkei*, and P. Ienne

FPGA'21, Online, 01.03.2021

École Polytechnique Fédérale de Lausanne
*IMEC

Metal Stack Evolution

N16: Wu et al.,

"A 16nm FinFET CMOS technology for mobile SoC and computing applications",

IEDM'13

N7: Wu et al.,

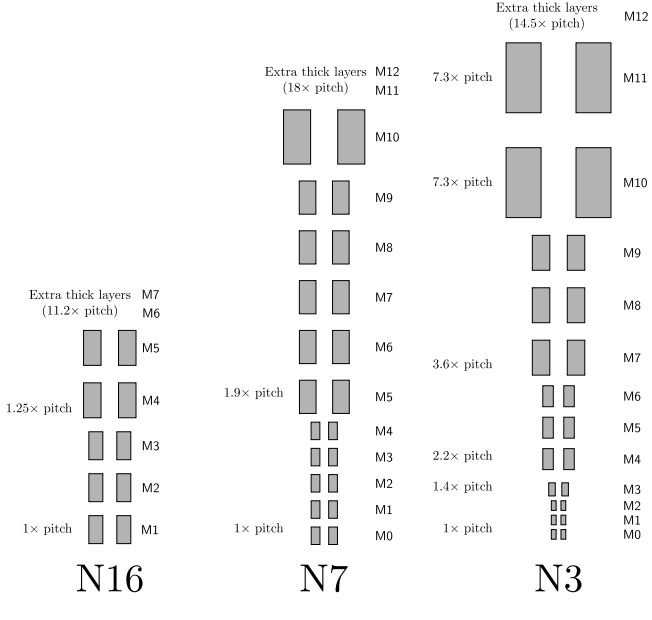
"A 7nm CMOS platform technology featuring 4th generation FinFET transistors with a 0.027 um2 high density 6-T SRAM cell for mobile SoC applications",

IEDM'16

N3: Prasad et al.,

"Buried power rails and back-side power grids: Arm® CPU power delivery network design beyond 5nm",

IEDM'19



Metal Stack Evolution

N16: Wu et al.,

"A 16nm FinFET CMOS technology for mobile SoC and computing applications",

IEDM'13

N7: Wu et al.,

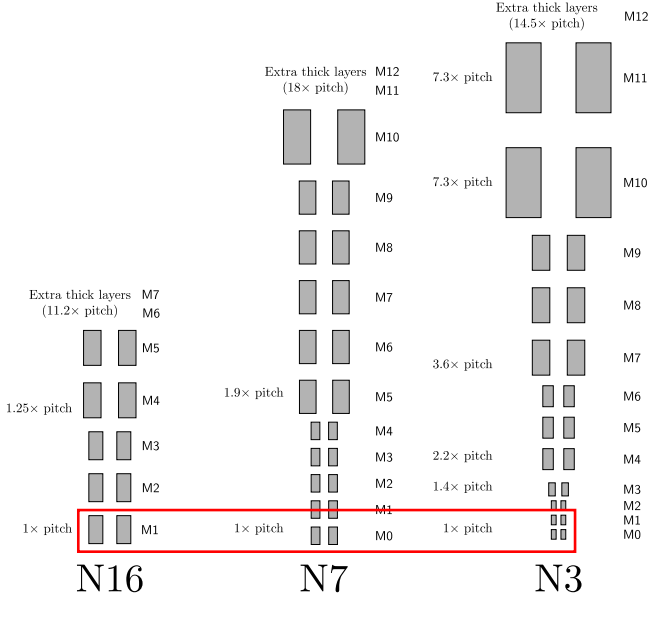
"A 7nm CMOS platform technology featuring 4th generation FinFET transistors with a 0.027 um2 high density 6-T SRAM cell for mobile SoC applications",

IEDM'16

N3: Prasad et al.,

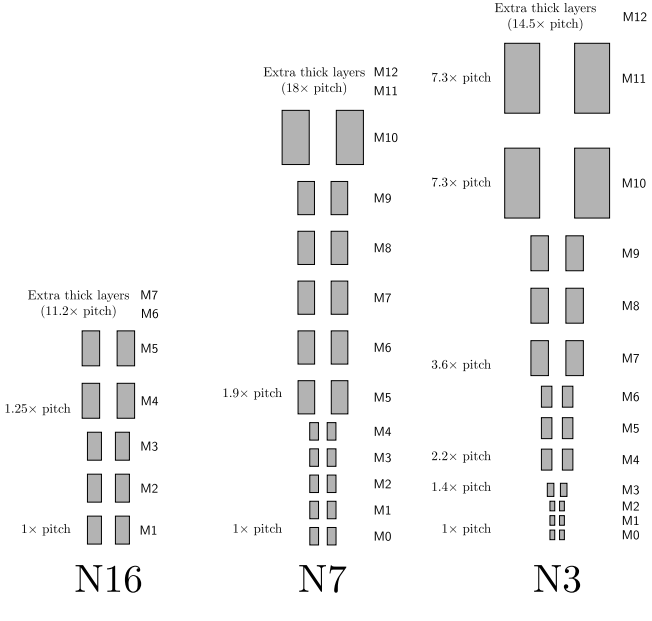
"Buried power rails and back-side power grids: Arm® CPU power delivery network design beyond 5nm",

IEDM'19



Some High School Physics

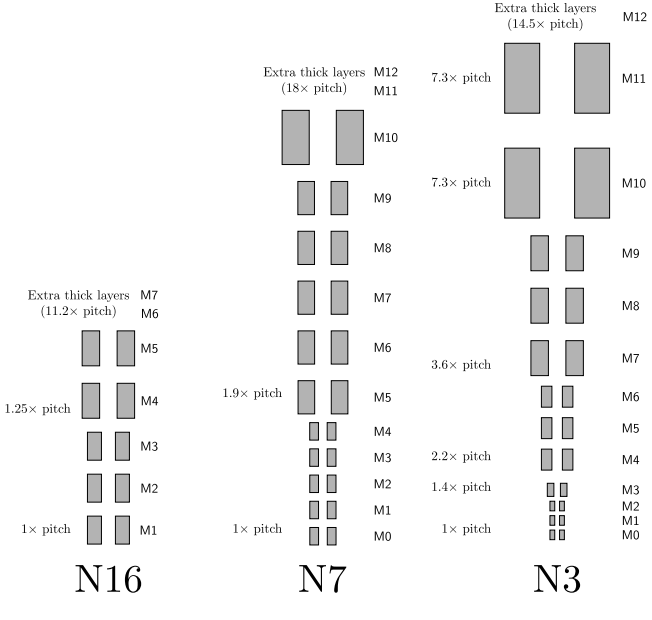
$$R = \frac{\rho l}{S}$$



Some High School Physics

$$R = \frac{\rho l}{S}$$

$$S \searrow \implies R \nearrow$$

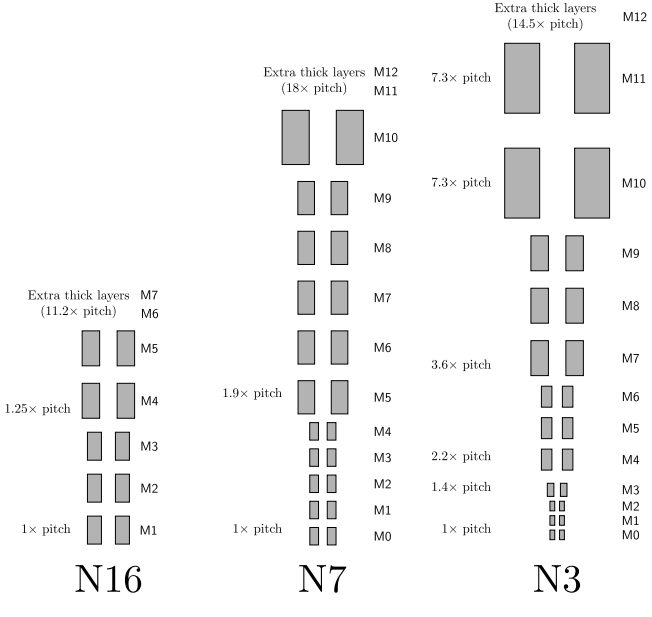


Some High School Physics

$$R = \frac{\rho l}{S}$$

$$S \searrow \implies R \nearrow$$

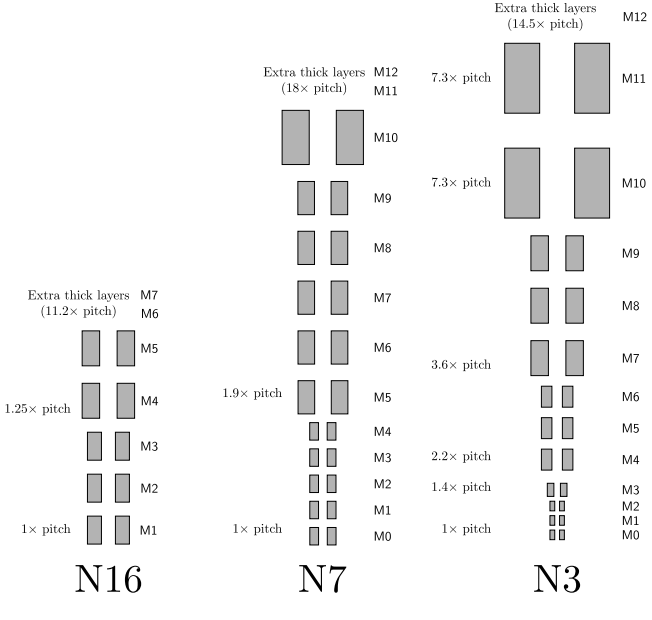
$$t_d = RC \ln 2$$



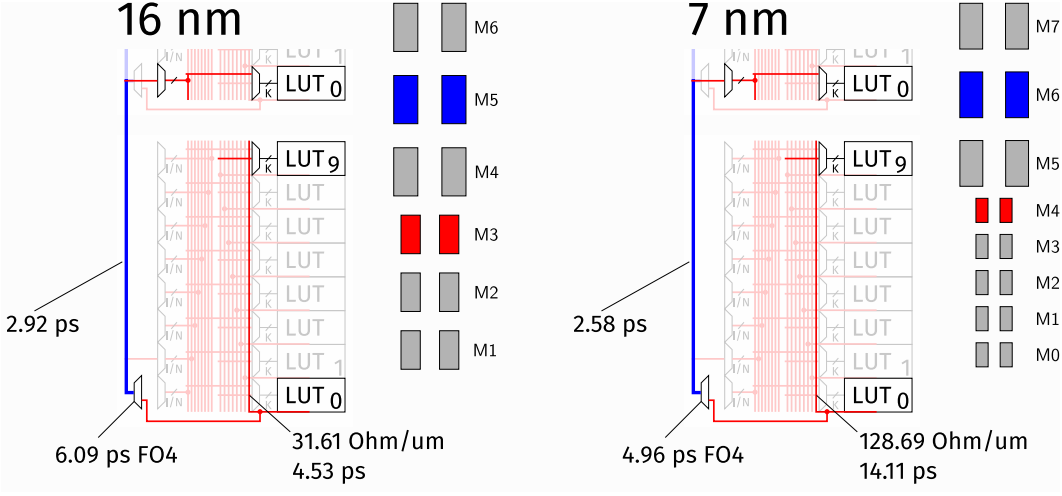
Some High School Physics

$$R = \frac{\rho l}{S}$$
$$S \searrow \implies R \nearrow$$

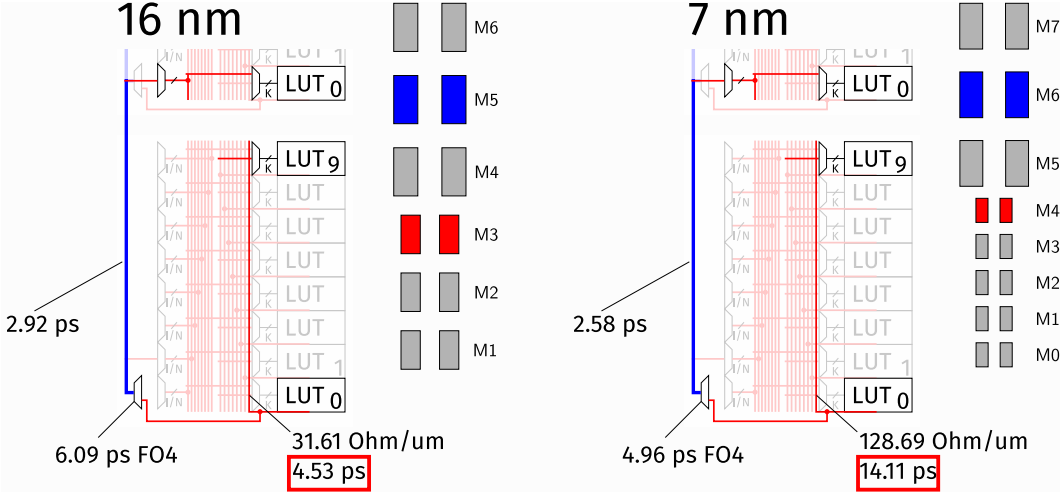
$$t_d = RC \ln 2$$
$$R \nearrow \implies t_d \nearrow$$



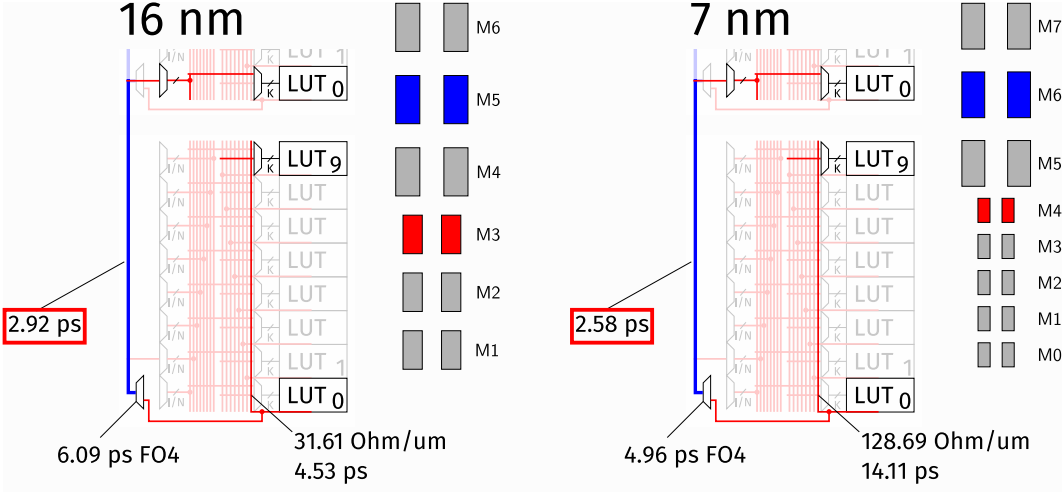
An Illustrative Example



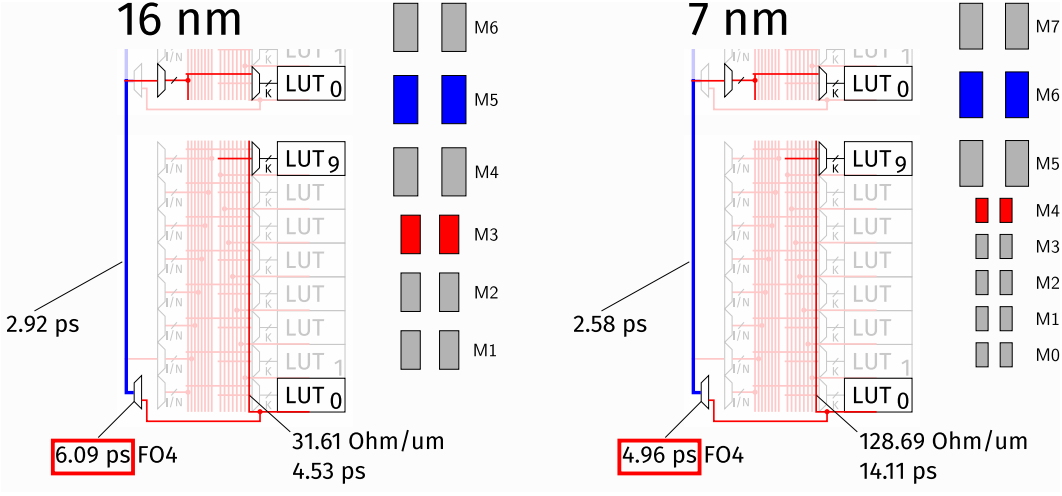
An Illustrative Example



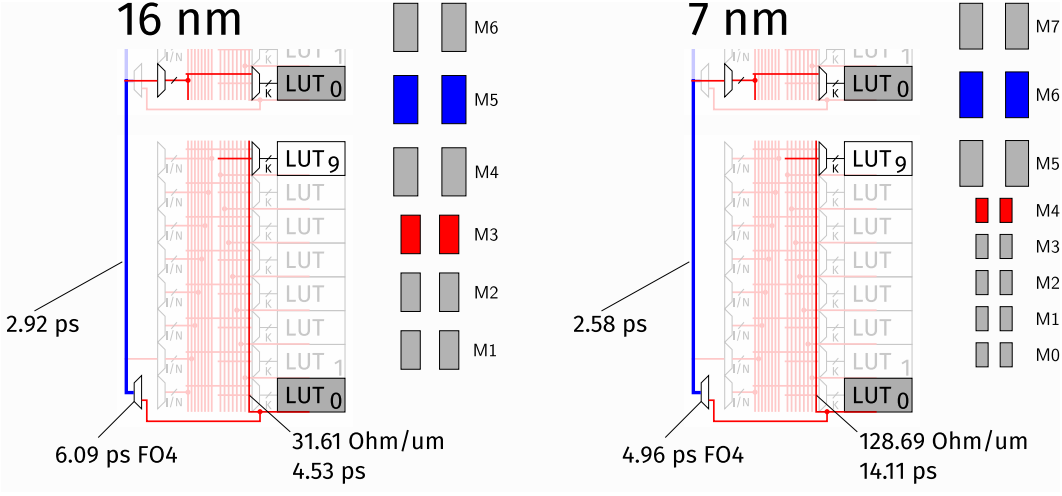
An Illustrative Example



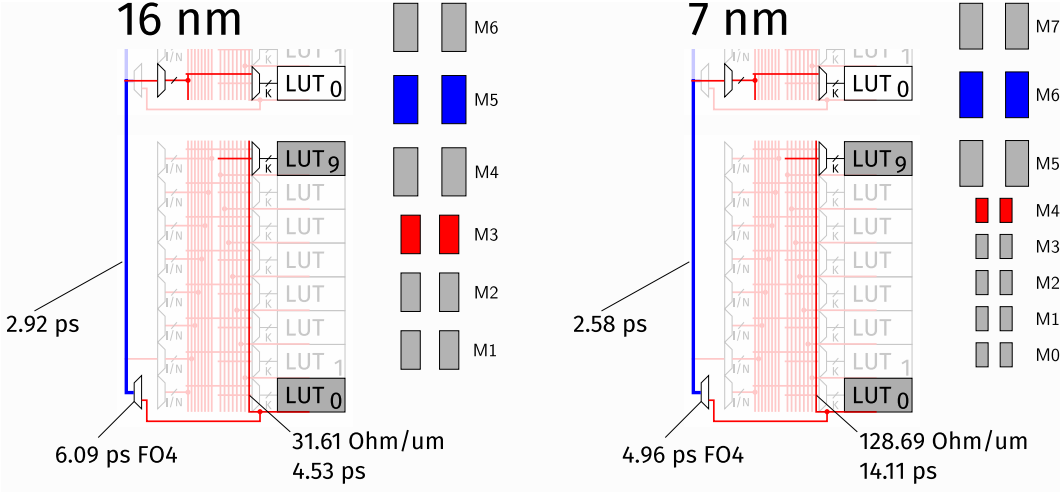
An Illustrative Example



An Illustrative Example



An Illustrative Example



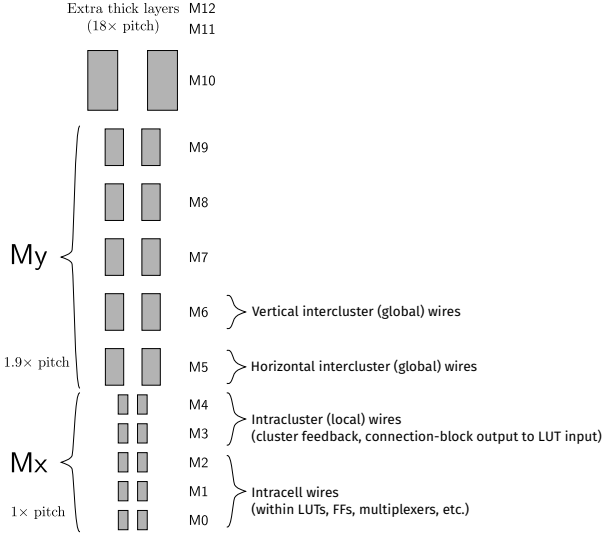
Outline

Metal Stack Modeling

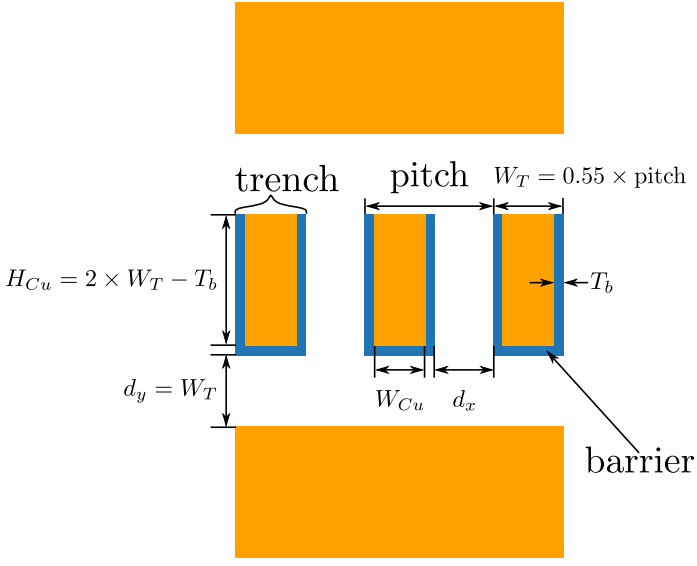
Layer Planning

Two pitch options considered:

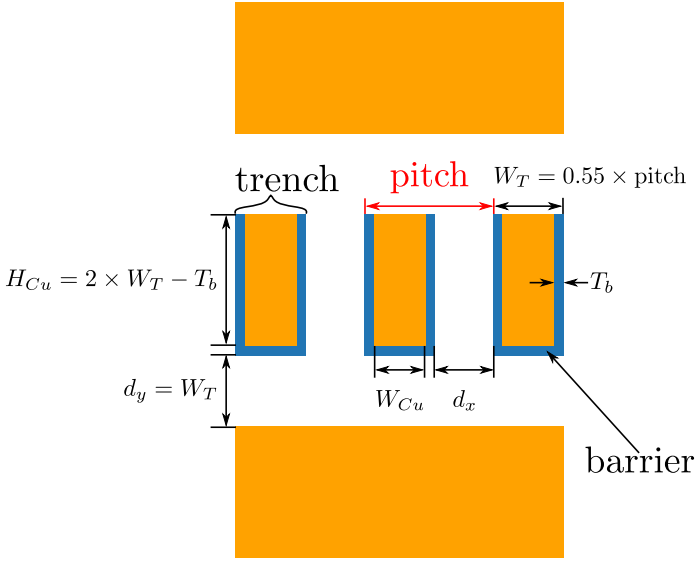
- Mx for intracluster (local) wires
- My for intercluster (global) wires



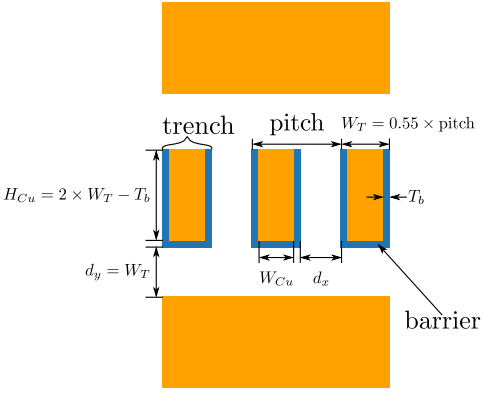
Wire Geometry



Wire Geometry



Resistance

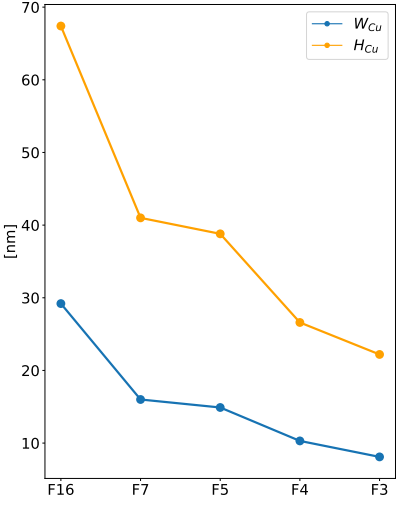


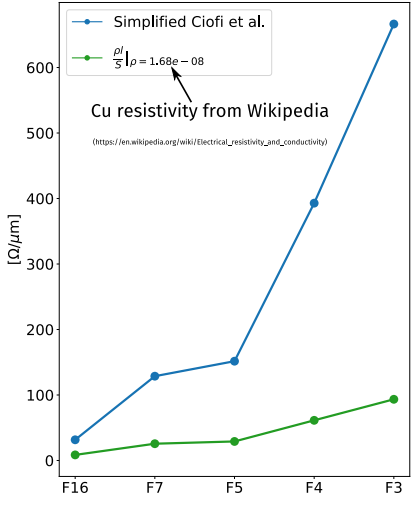
Ciofi et al., “Impact of Wire Geometry on Interconnect RC and Circuit Delay”, T-ED, 2016

$$R' = \frac{1}{H_{Cu} W_{Cu}} \left(32.05 + 615 \left(\frac{\tanh(0.133 W_{Cu})}{W_{Cu}} + \frac{\tanh(0.133 H_{Cu})}{H_{Cu}} \right) \right) \quad (1)$$

Resistance: Mx-Wires

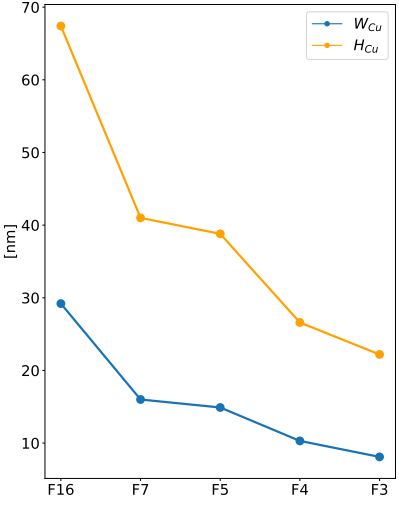
	F16	F7	F5	F4	F3
pitch [nm]	64	40	38	26	22

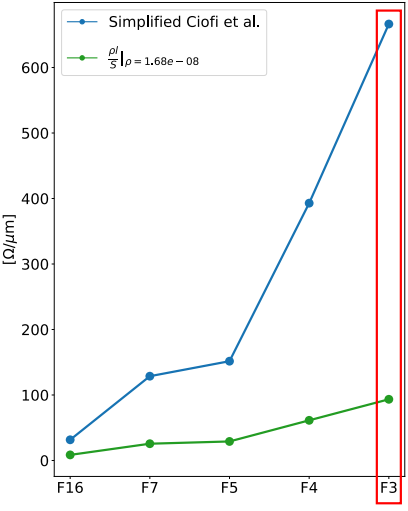




Resistance: Mx-Wires

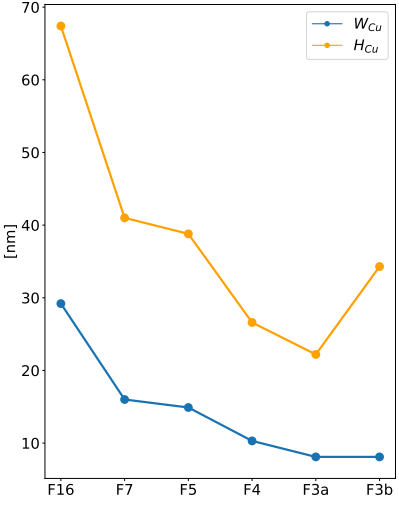
	F16	F7	F5	F4	F3
pitch [nm]	64	40	38	26	22

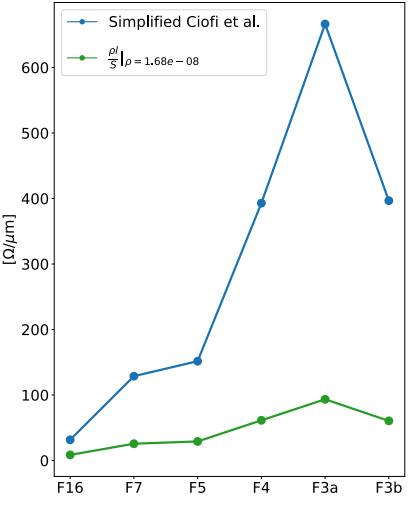




Resistance: Mx-Wires

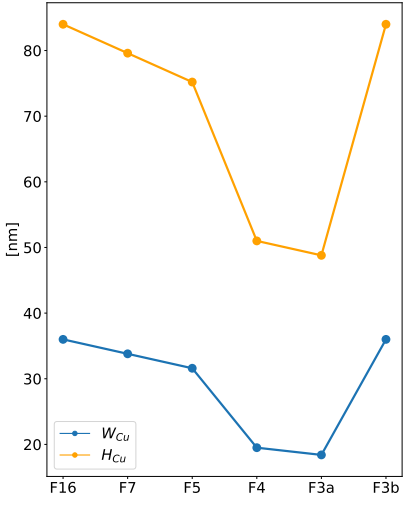
	F16	F7	F5	F4	F3a	F3b
pitch [nm]	64	40	38	26	22	22

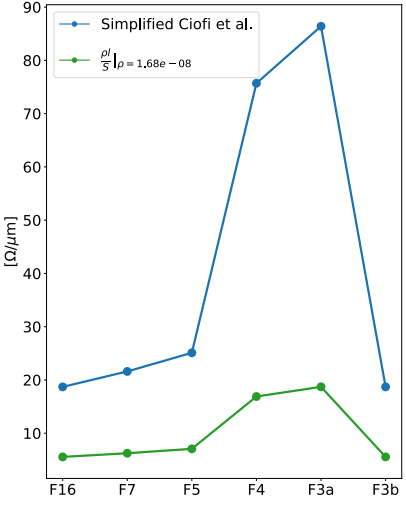


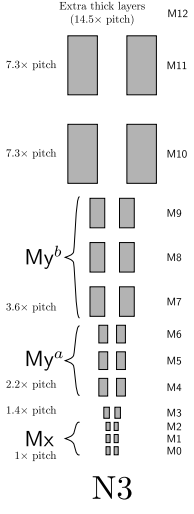


Resistance: My-Wires

	F16	F7	F5	F4	F3a	F3b
pitch [nm]	80	76	72	50	48	80



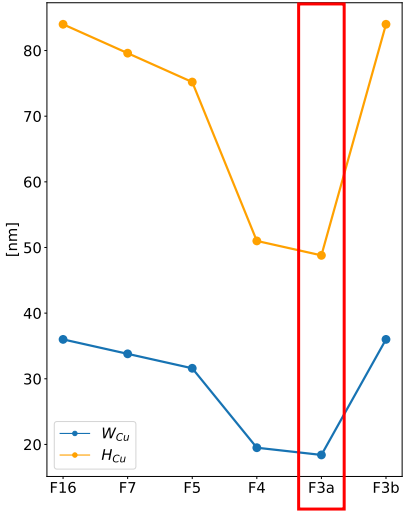


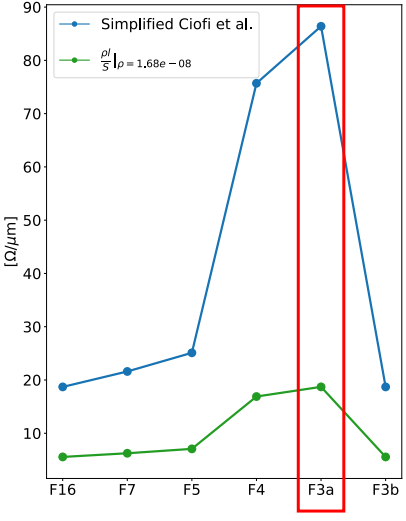


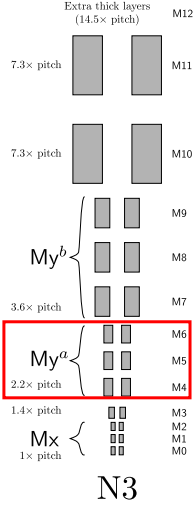
N3

Resistance: My-Wires

	F16	F7	F5	F4	F3a	F3b
pitch [nm]	80	76	72	50	48	80



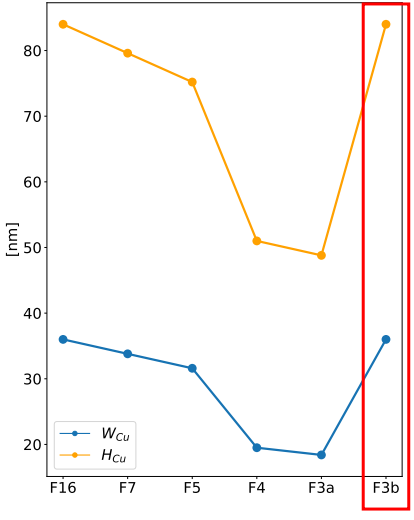


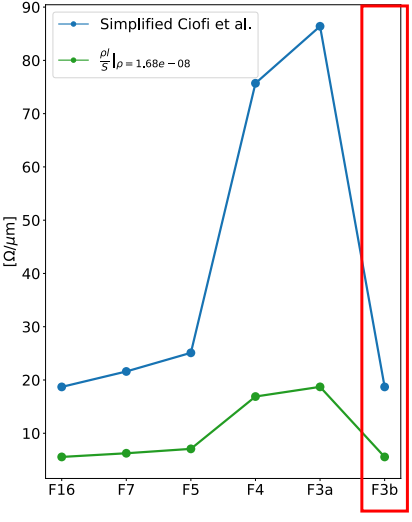


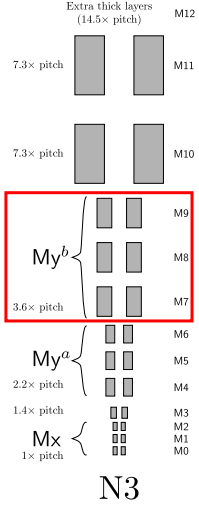
N3

Resistance: My-Wires

	F16	F7	F5	F4	F3a	F3b
pitch [nm]	80	76	72	50	48	80







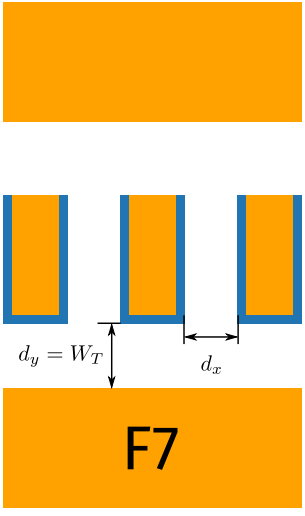
N3

Capacitance

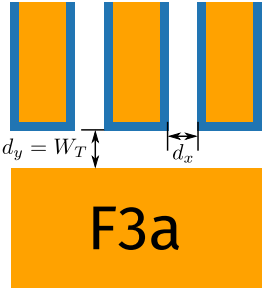
- Capacitance is less sensitive to scaling than resistance

Capacitance

- Capacitance is less sensitive to scaling than resistance

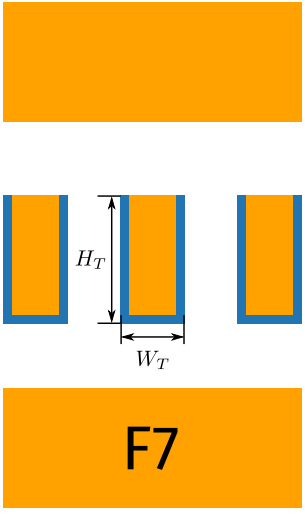


$$C = K\varepsilon_0 \frac{S}{d}$$
$$d \nearrow \implies C \nearrow$$

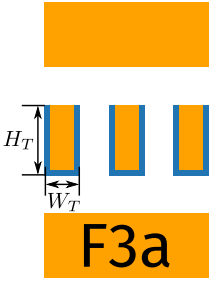


Capacitance

- Capacitance is less sensitive to scaling than resistance



$$C = K\varepsilon_0 \frac{S}{d}$$
$$d \nearrow \implies C \nearrow$$
$$S \searrow \implies C \searrow$$



Capacitance

Wong et al., “Modeling of interconnect capacitance, delay, and crosstalk in VLSI”, T-SM, 2000

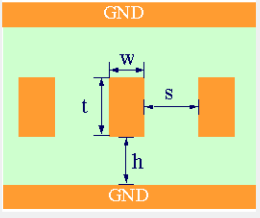
Predictive Technology Model

- Introduction
- Latest Models
- Nano-CMOS
- Post-Silicon
- Interconnect

INTERCONNECT

Structure 2

Coupling lines between two metal ground planes (for local and intermediate layers)



Predictive Technology Model (PTM), Nanoscale Integration and Modeling Group, Arizona State University (ptm.asu.edu)

Area and Wirelength Modeling

Purpose

Good area and length models are necessary for

- Delay measurement

Purpose

Good area and length models are necessary for

- Delay measurement
- Determining the maximum number of tracks in the routing channels

Purpose

Good area and length models are necessary for

- Delay measurement
- Determining the maximum number of tracks in the routing channels

Typical models based on transistor counting
are insufficient for scaled technologies

Purpose

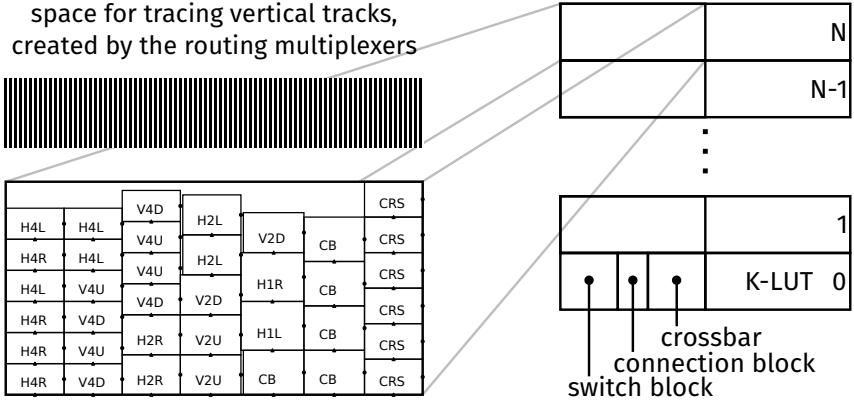
Good area and length models are necessary for

- Delay measurement
- Determining the maximum number of tracks in the routing channels

Typical models based on transistor counting
are insufficient for scaled technologies

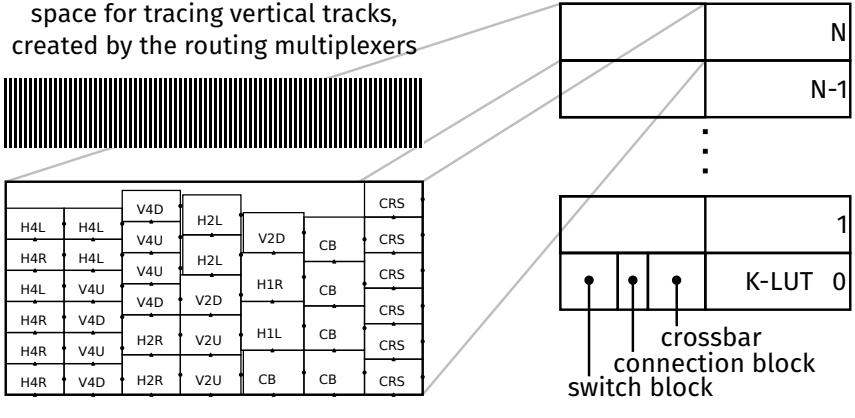
Khan and Ye, “An Evaluation on the Accuracy of the Minimum Width Transistor Area Models in Ranking the Actual Layout Area of FPGA”, FPL’16

Floorplan



Lewis et al., “Architectural enhancements in Stratix V”, FPGA’13

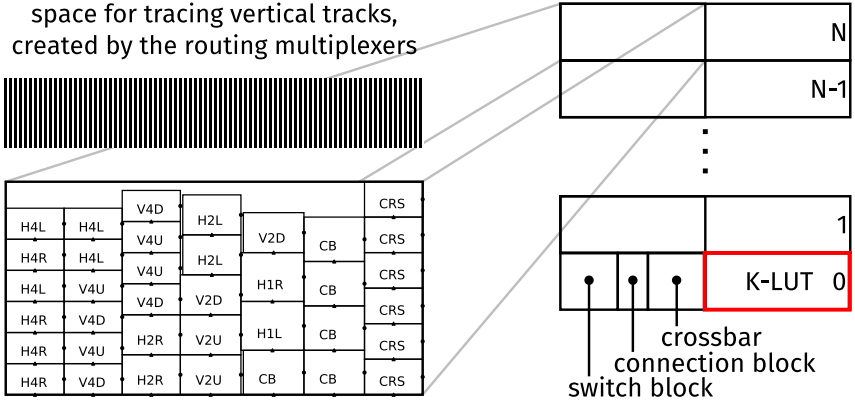
Floorplan



Lewis et al., “Architectural enhancements in Stratix V”, FPGA’13

Chromczak et al., “Architectural enhancements in Intel Agilex FPGAs”, FPGA’20

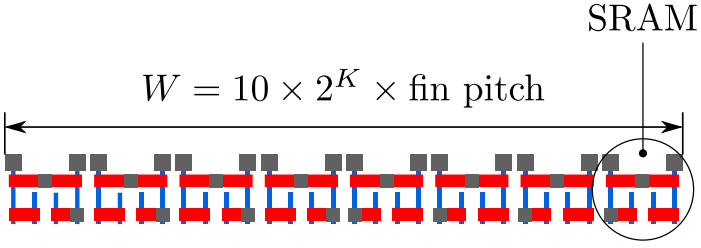
Floorplan



Lewis et al., “Architectural enhancements in Stratix V”, FPGA’13

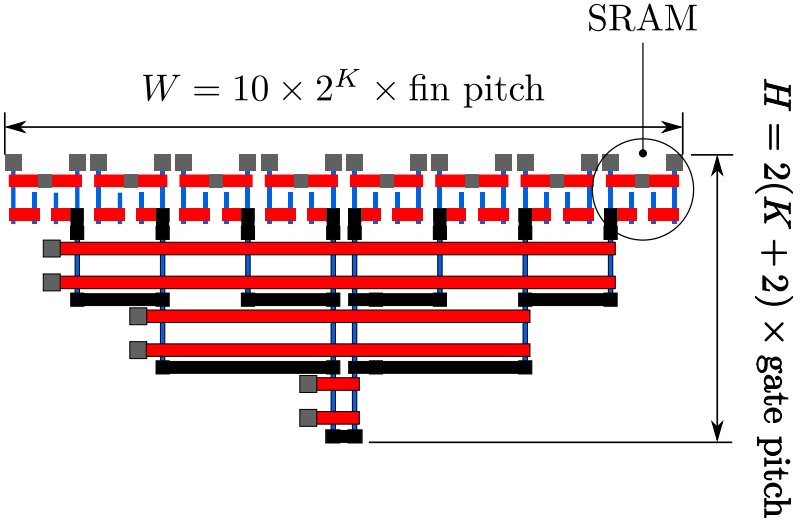
Chromczak et al., “Architectural enhancements in Intel Agilex FPGAs”, FPGA’20

LUTs



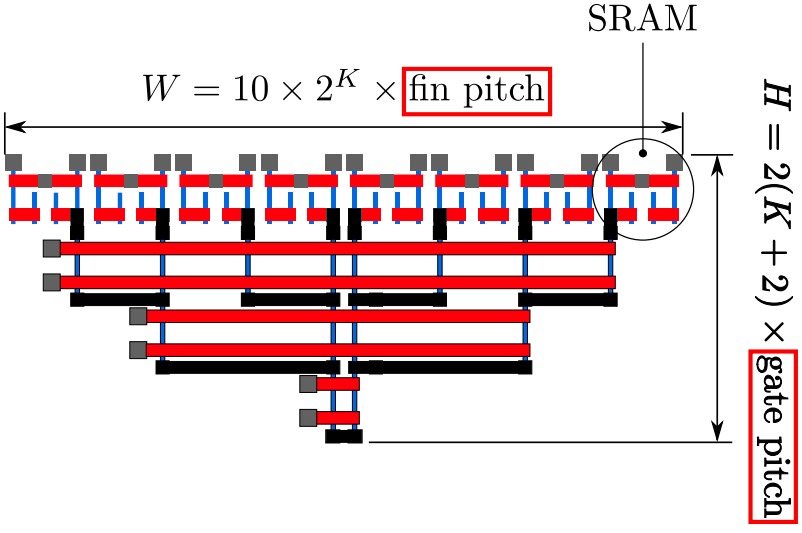
Abusultan and Khatri, “A comparison of FinFET-based FPGA LUT designs”, GLSVLSI’14

LUTs



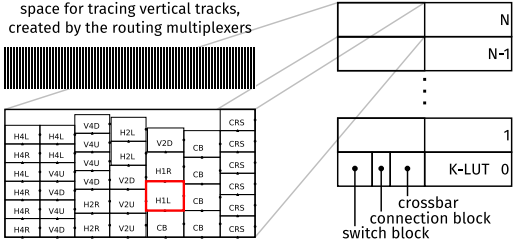
Abusultan and Khatri, “A comparison of FinFET-based FPGA LUT designs”, GLSVLSI’14

LUTs

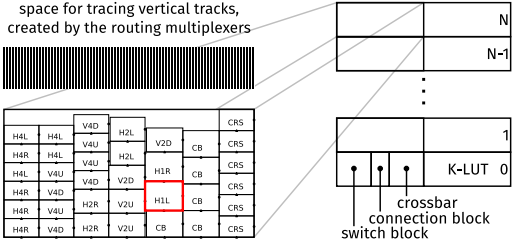


Abusultan and Khatri, “A comparison of FinFET-based FPGA LUT designs”, GLSVLSI’14

Multiplexers



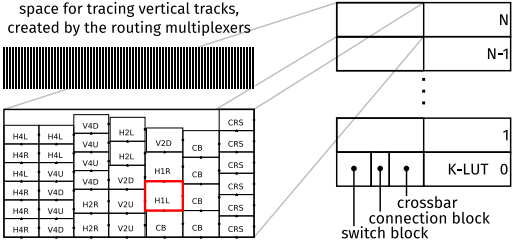
Multiplexers



- All muxes transmission-gate-based

  Chromczack et al., FPGA'20

Multiplexers



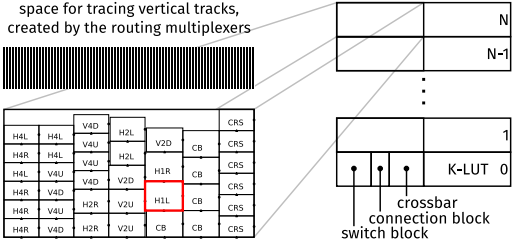
- All muxes transmission-gate-based

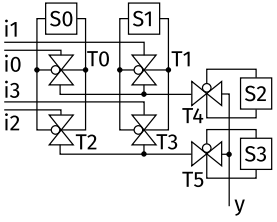
  Chromczack et al., FPGA'20

- All transmission-gates of minimum drive-strength (1 fin)

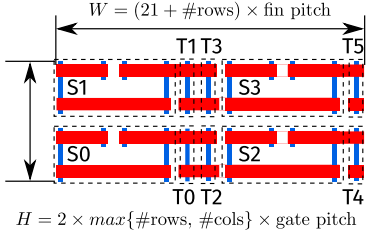
  Chiasson, MSc Thesis, University of Toronto, 2013

Multiplexers

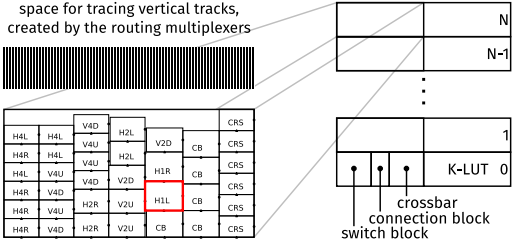


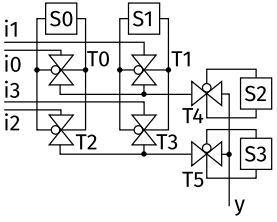


- All muxes transmission-gate-based
  Chromczack et al., FPGA'20
- All transmission-gates of minimum drive-strength (1 fin)
  Chiasson, MSc Thesis, University of Toronto, 2013

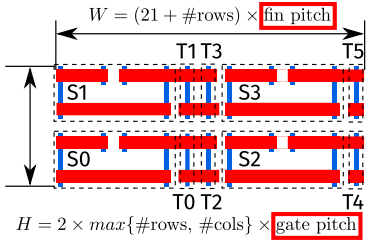


Multiplexers



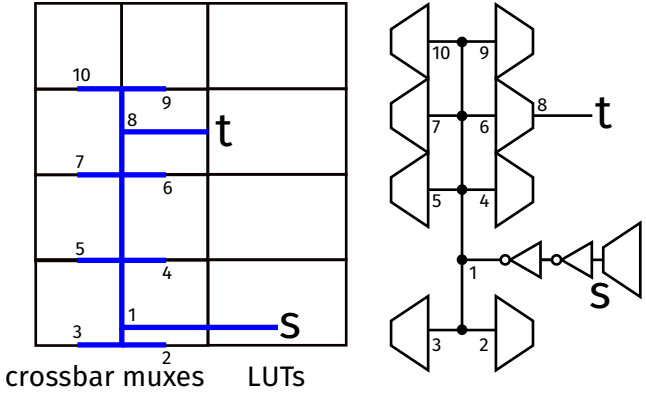


- All muxes transmission-gate-based
  Chromczack et al., FPGA'20
- All transmission-gates of minimum drive-strength (1 fin)
  Chiasson, MSc Thesis, University of Toronto, 2013

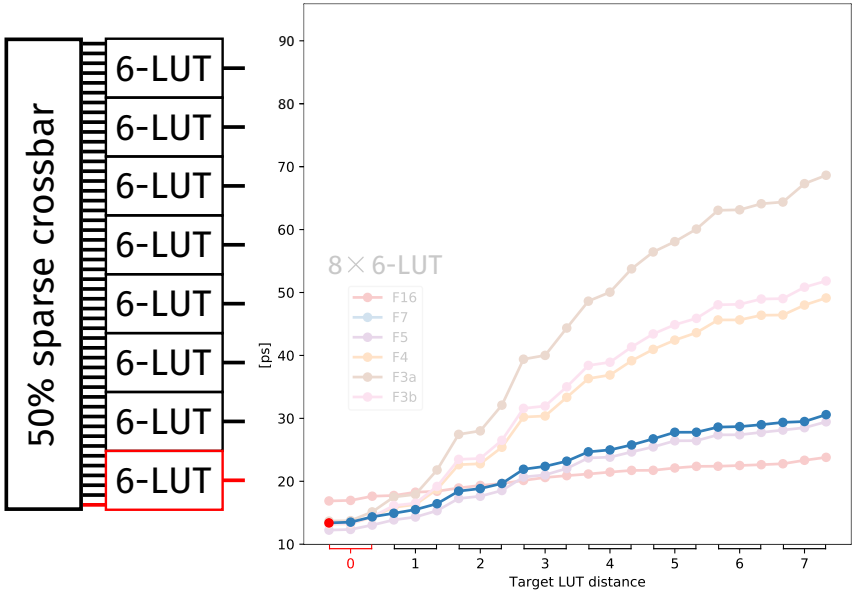


Delay Measurement

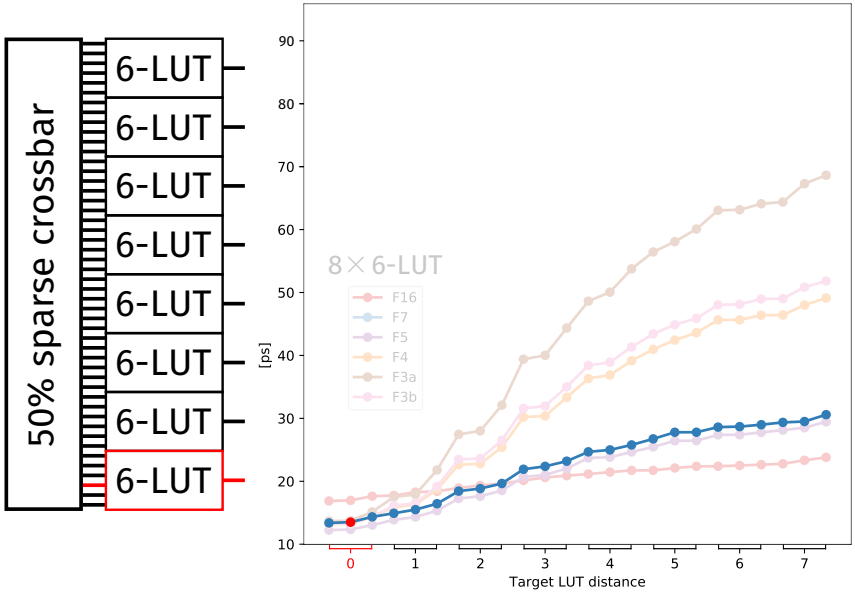
Local Connections



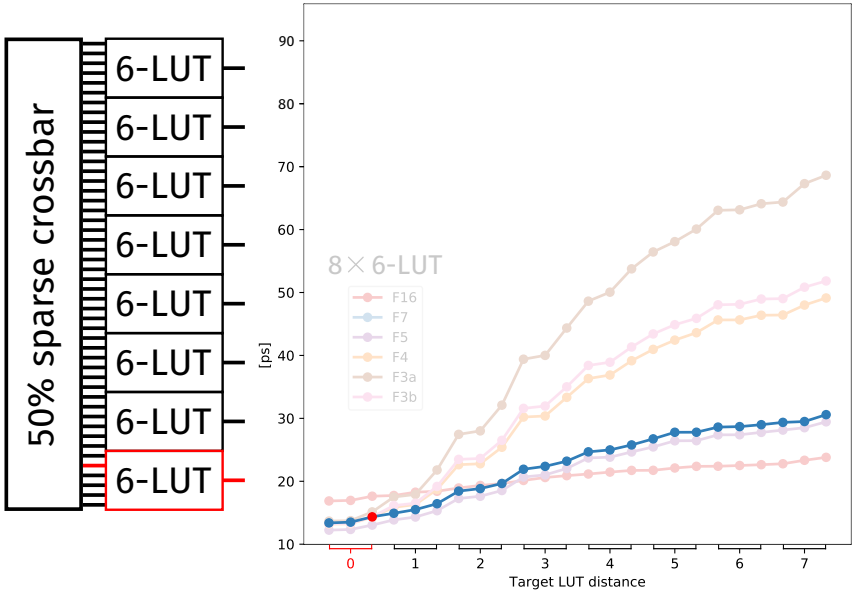
Local Connections: Cluster Feedback Delays



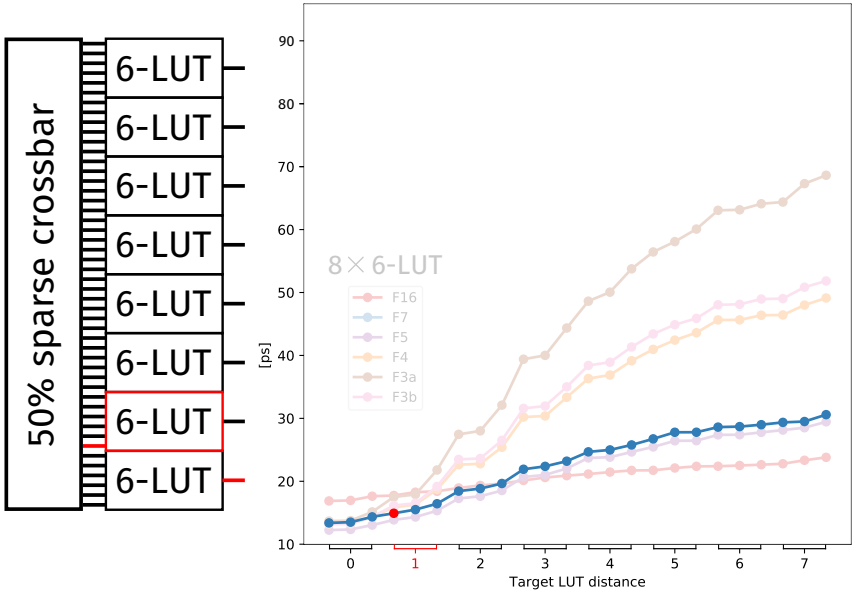
Local Connections: Cluster Feedback Delays



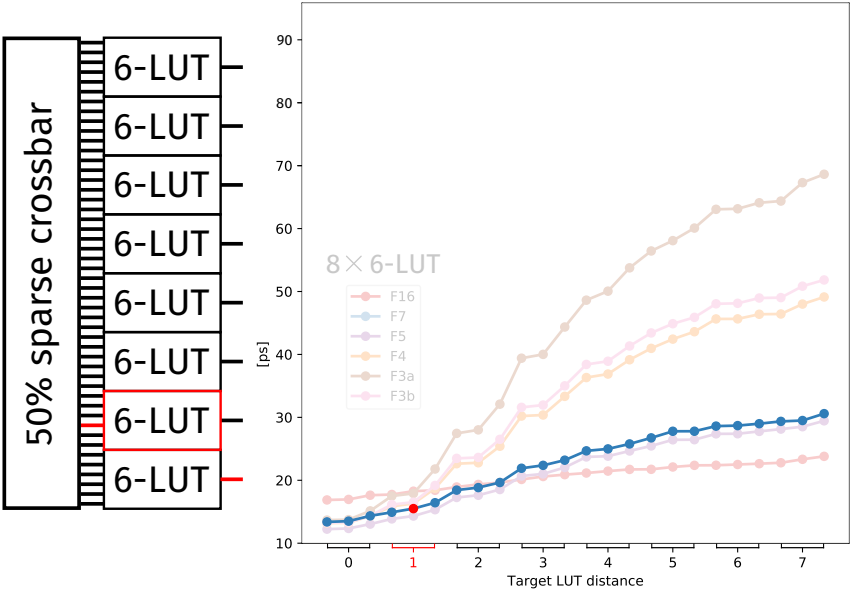
Local Connections: Cluster Feedback Delays



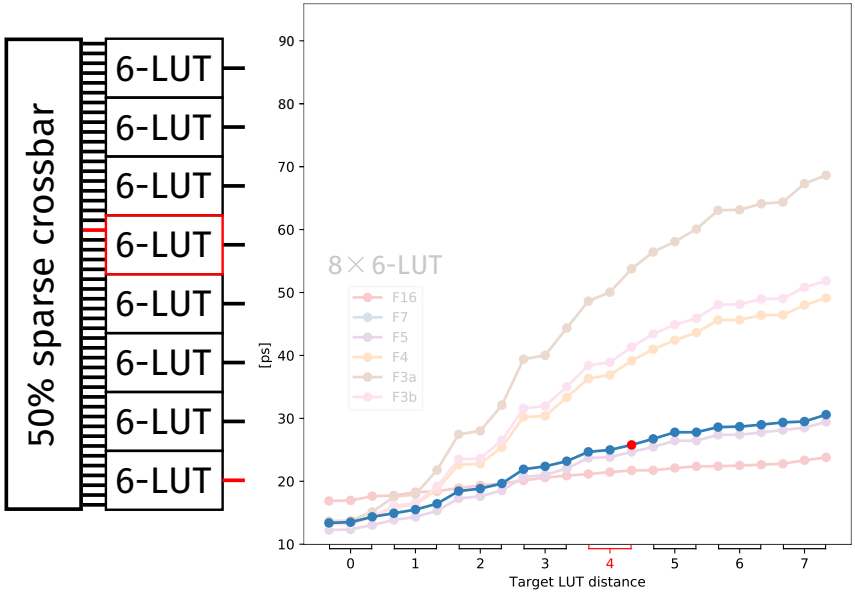
Local Connections: Cluster Feedback Delays



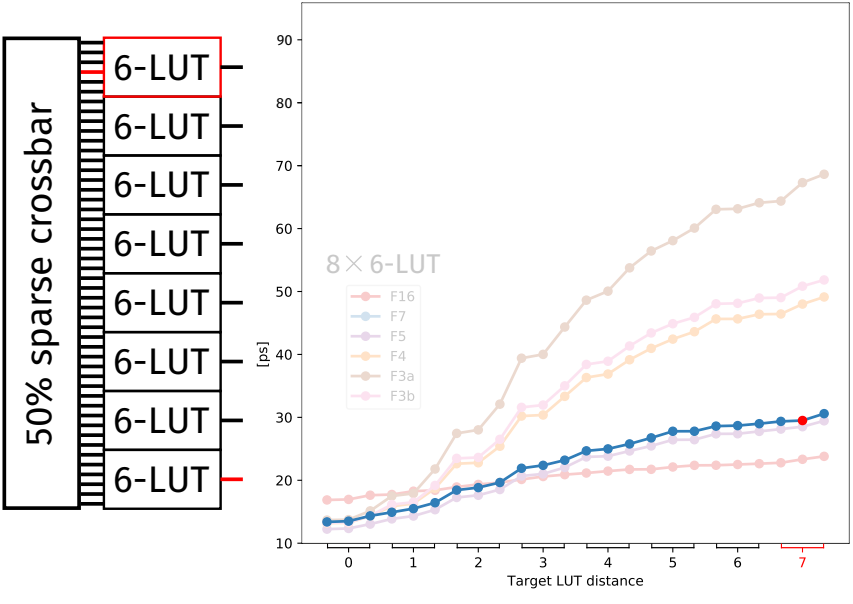
Local Connections: Cluster Feedback Delays



Local Connections: Cluster Feedback Delays

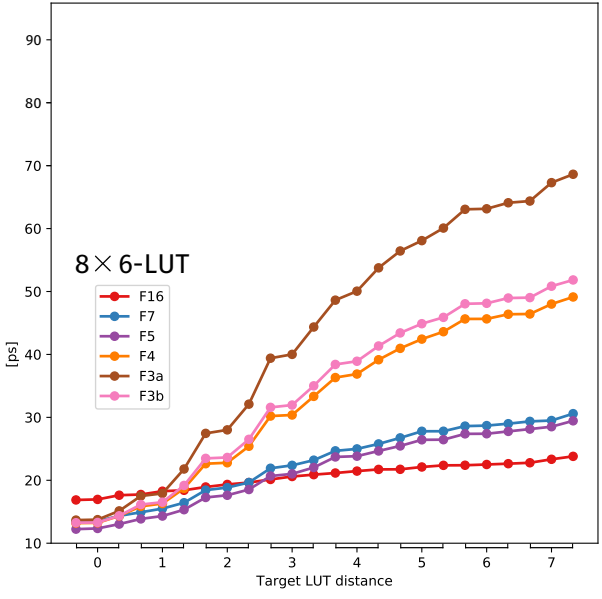


Local Connections: Cluster Feedback Delays



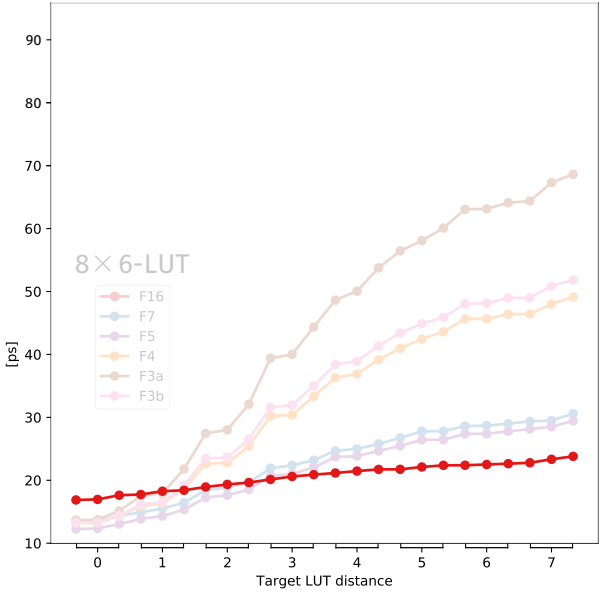
Local Connections: Cluster Feedback Delays

	pitch [nm]	R' [Ω/μm]
F16	64	31.6
F7	40	128.7
F5	38	151.6
F4	26	392.9
F3a	22	666.4
F3b	22	396.7



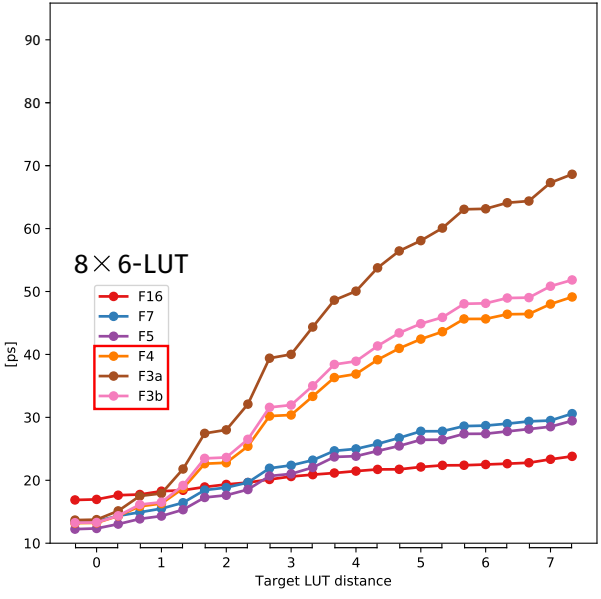
Local Connections: Cluster Feedback Delays

	pitch [nm]	R' [Ω/μm]
F16	64	31.6
F7	40	128.7
F5	38	151.6
F4	26	392.9
F3a	22	666.4
F3b	22	396.7



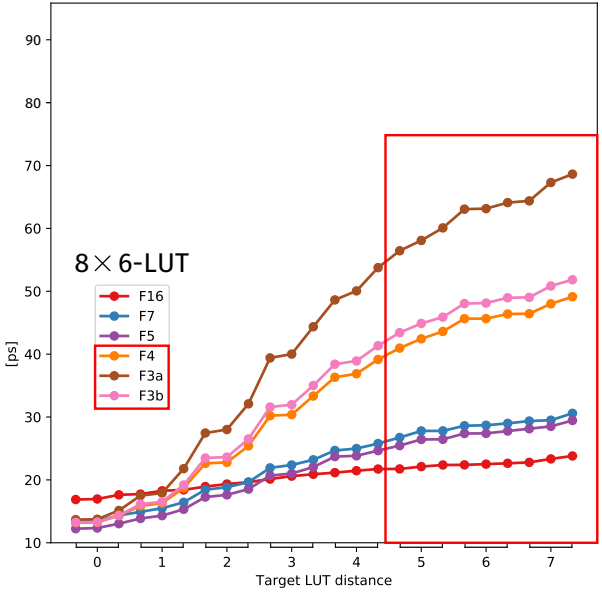
Local Connections: Cluster Feedback Delays

	pitch [nm]	R' [Ω/μm]
F16	64	31.6
F7	40	128.7
F5	38	151.6
F4	26	392.9
F3a	22	666.4
F3b	22	396.7



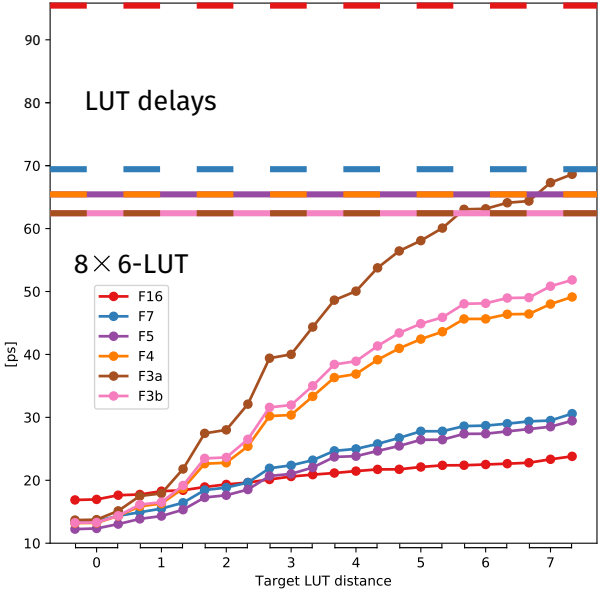
Local Connections: Cluster Feedback Delays

	pitch [nm]	R' [Ω/μm]
F16	64	31.6
F7	40	128.7
F5	38	151.6
F4	26	392.9
F3a	22	666.4
F3b	22	396.7



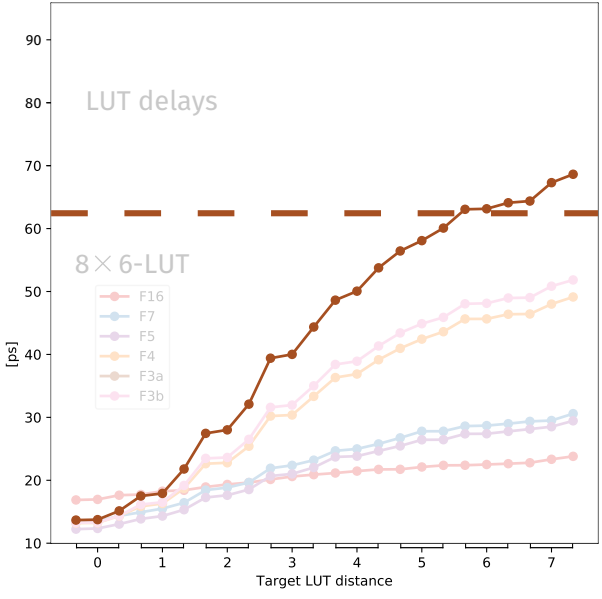
Local Connections: Cluster Feedback Delays

	pitch [nm]	R' [Ω/μm]
F16	64	31.6
F7	40	128.7
F5	38	151.6
F4	26	392.9
F3a	22	666.4
F3b	22	396.7

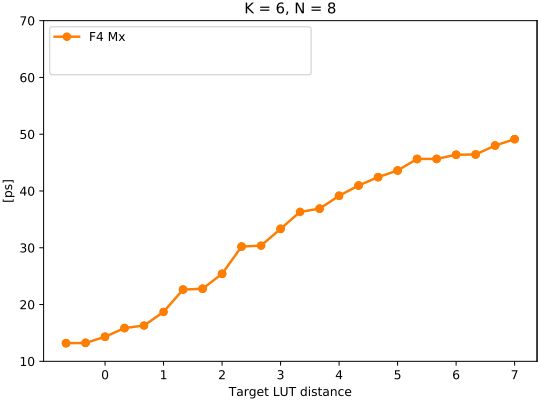


Local Connections: Cluster Feedback Delays

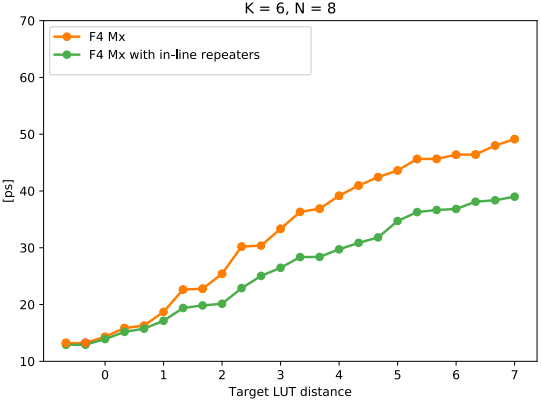
	pitch [nm]	R' [Ω/μm]
F16	64	31.6
F7	40	128.7
F5	38	151.6
F4	26	392.9
F3a	22	666.4
F3b	22	396.7

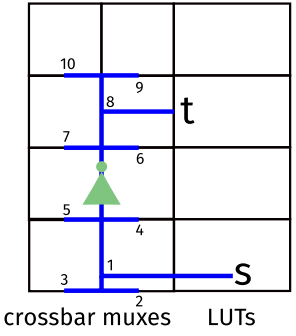


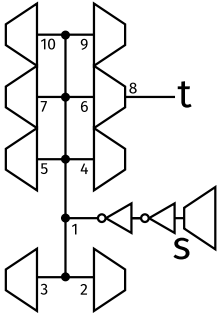
Local Connections: Cluster Feedback Delays



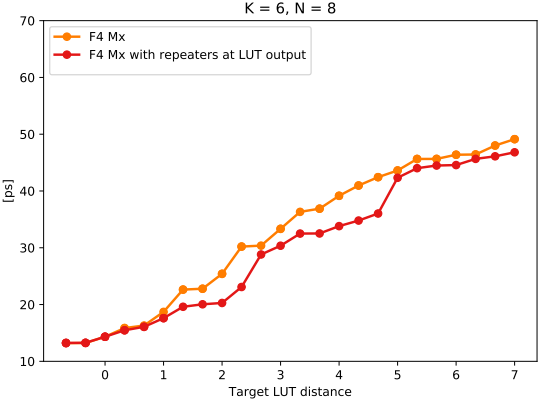
Local Connections: Cluster Feedback Delays

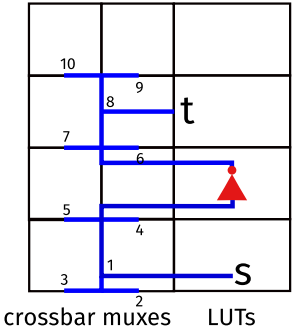


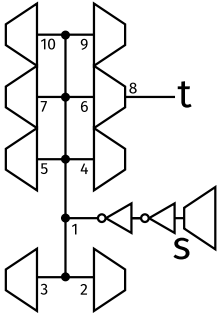




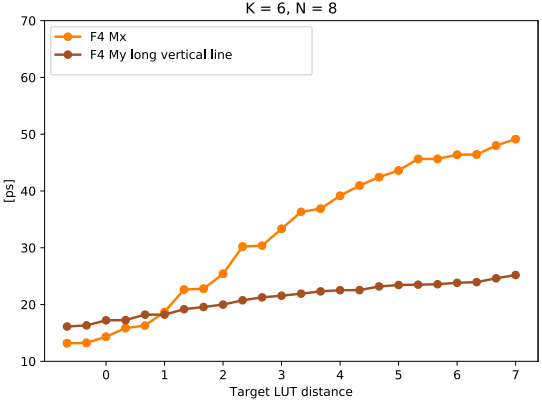
Local Connections: Cluster Feedback Delays

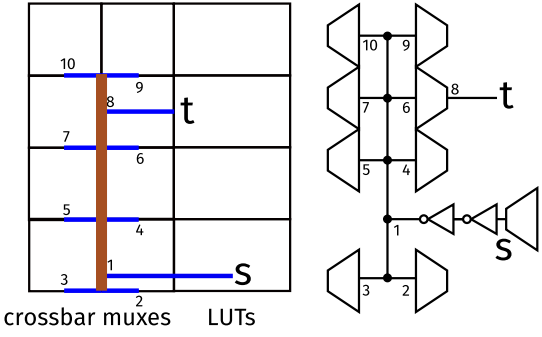




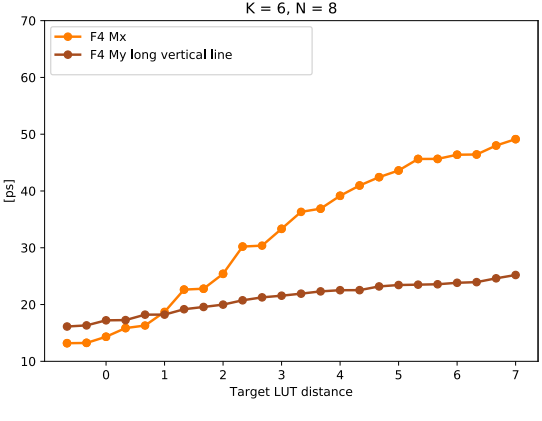


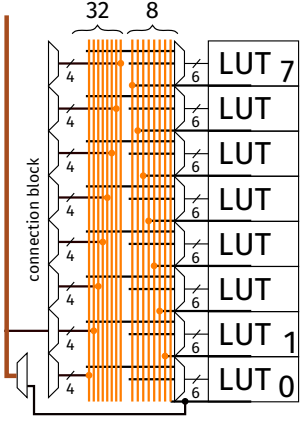
Local Connections: Cluster Feedback Delays



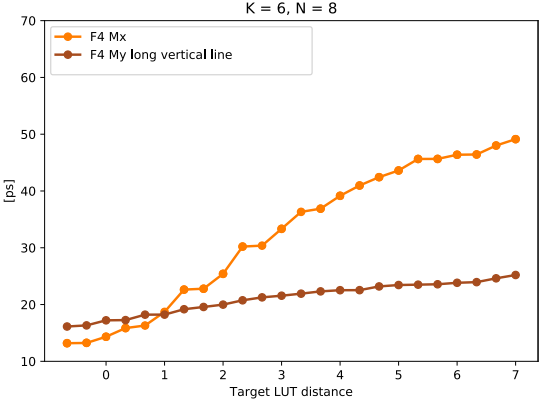


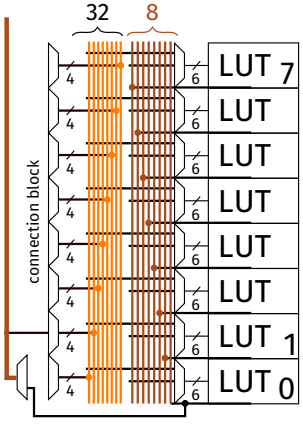
Thick Local Connections: Thick Metal is Scarce



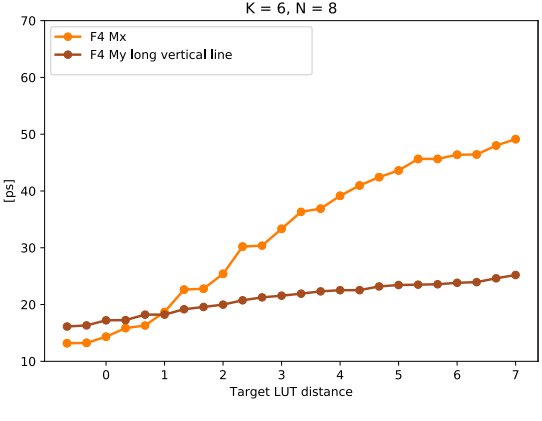


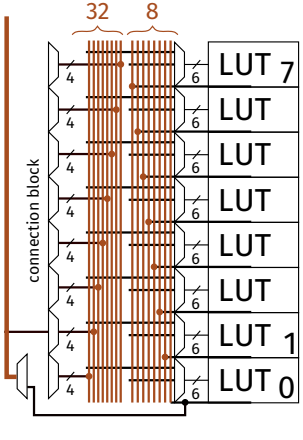
Thick Local Connections: Thick Metal is Scarce



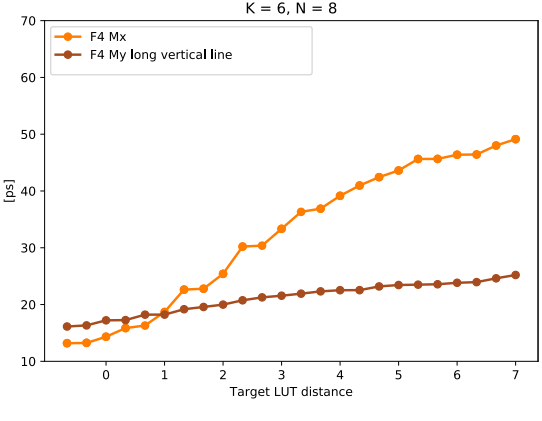


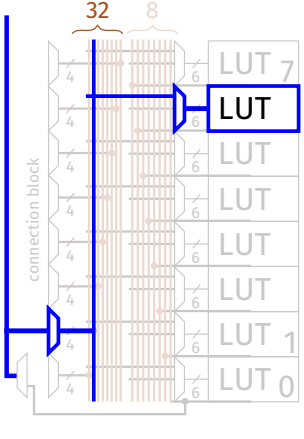
Thick Local Connections: Thick Metal is Scarce



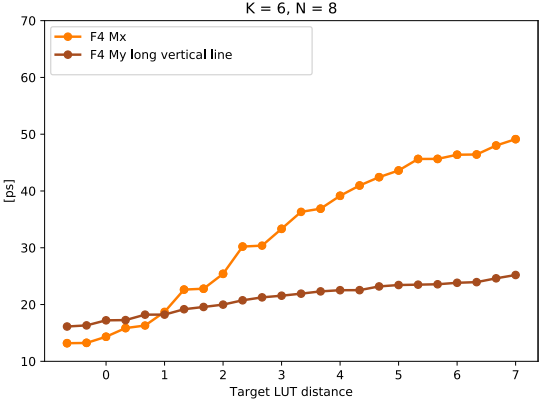


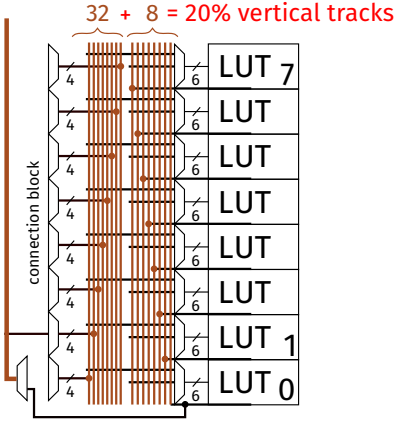
Thick Local Connections: Thick Metal is Scarce



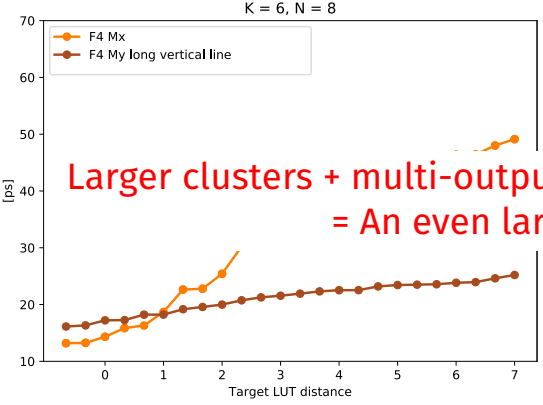


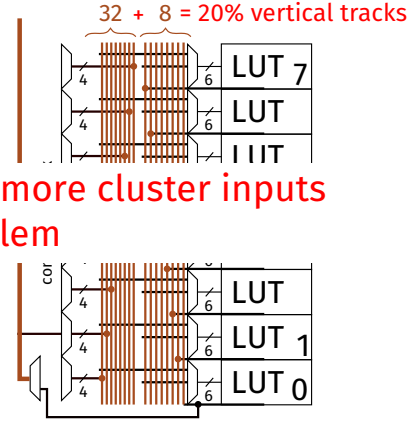
Thick Local Connections: Thick Metal is Scarce





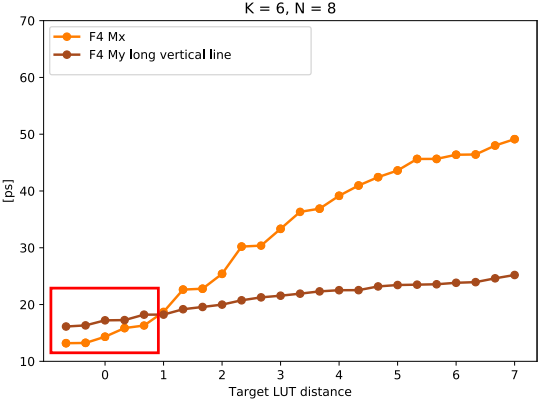
Thick Local Connections: Thick Metal is Scarce



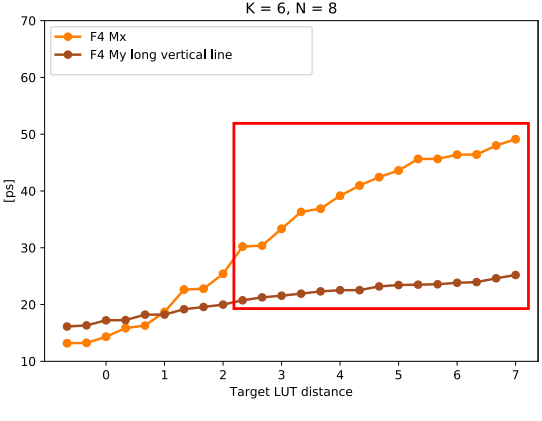


Larger clusters + multi-output LUTs + more cluster inputs
= An even larger problem

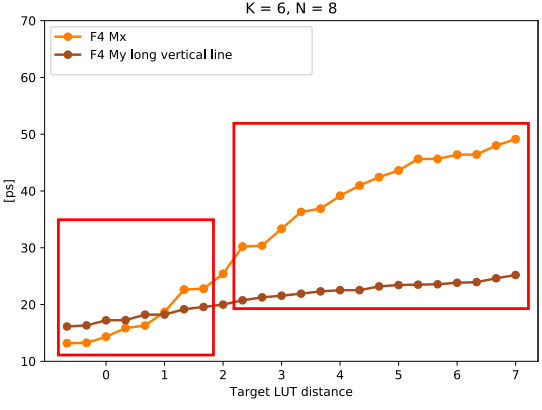
Thick Local Connections: Small Clusters to the Rescue

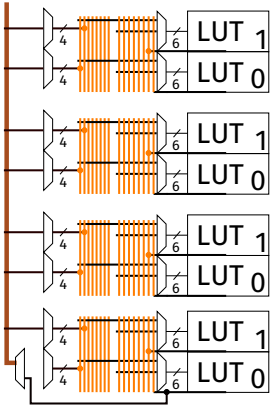


Thick Local Connections: Small Clusters to the Rescue

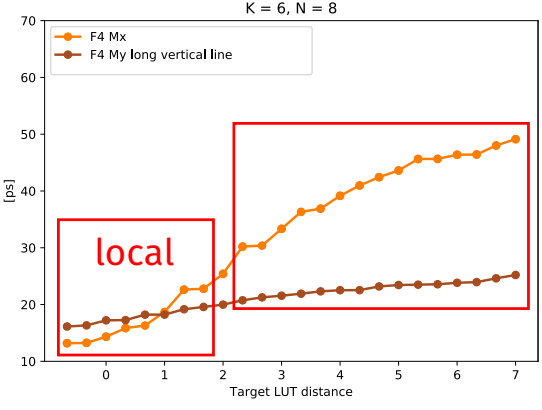


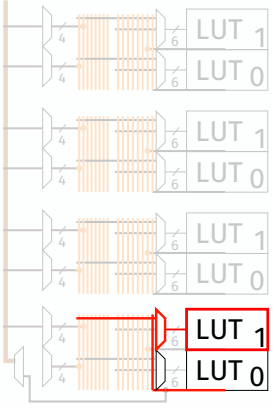
Thick Local Connections: Small Clusters to the Rescue



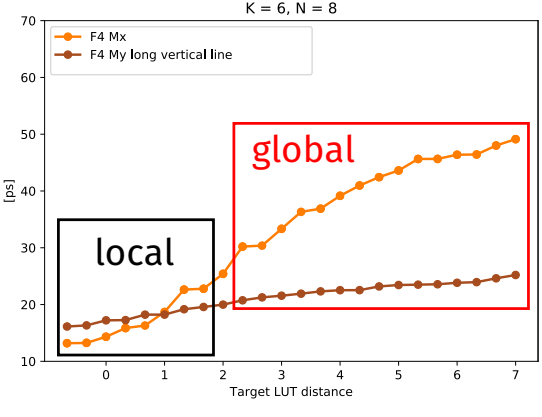


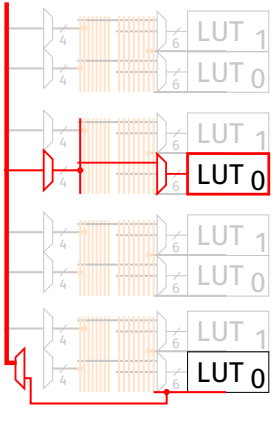
Thick Local Connections: Small Clusters to the Rescue



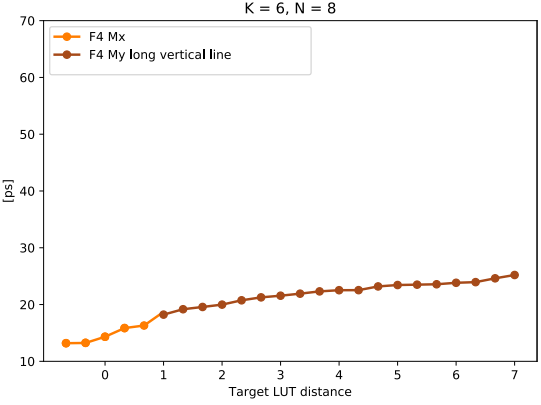


Thick Local Connections: Small Clusters to the Rescue





Thick Local Connections: Small Clusters to the Rescue



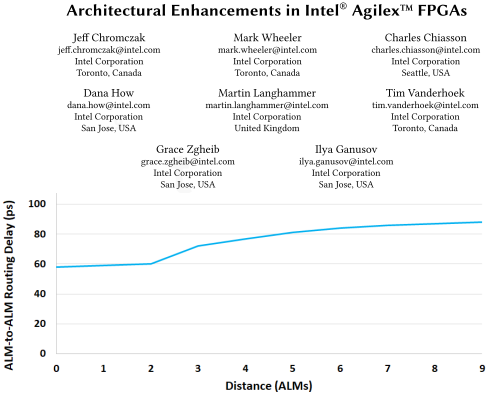
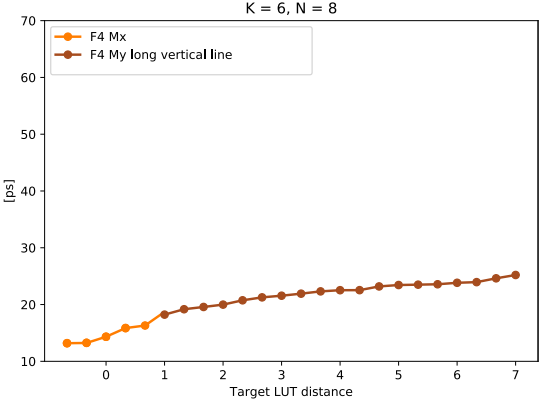
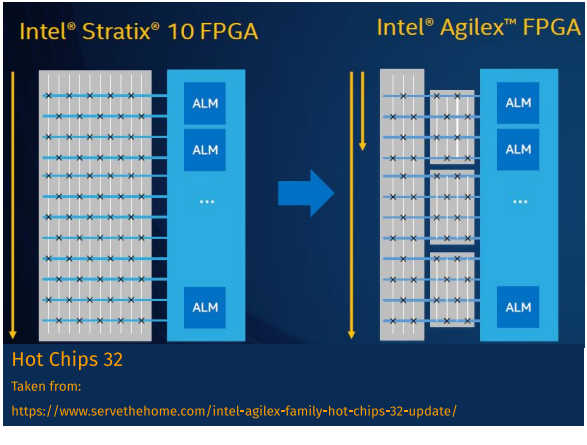


Figure 9: ALM-to-ALM routing delay improvement

Thick Local Connections: Small Clusters to the Rescue





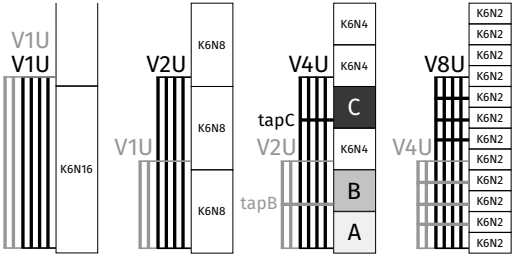
Exploring Cluster Sizes across Technology Nodes

Experimental Setup

- Clusters of 2, 4, 8, and 16 6-LUTs

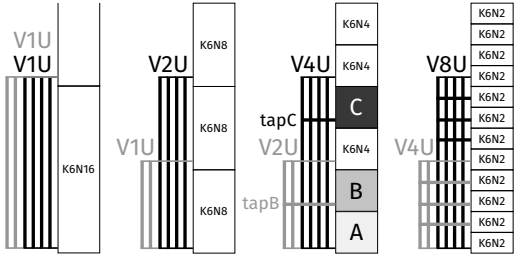
Experimental Setup

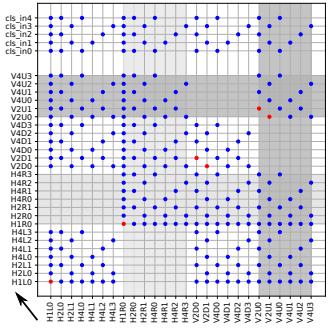
- Clusters of 2, 4, 8, and 16 6-LUTs
- Channel composition exploration



Experimental Setup

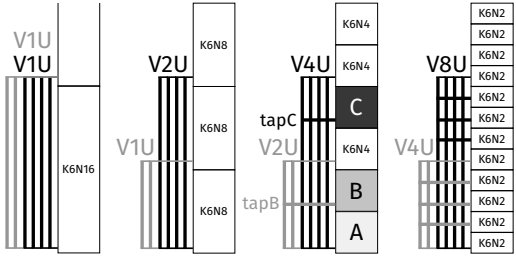
- Clusters of 2, 4, 8, and 16 6-LUTs
- Channel composition exploration
- Switch-patterns tailored for high-resistance lower metal

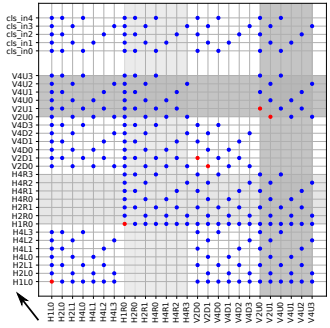




Experimental Setup

- Clusters of 2, 4, 8, and 16 6-LUTs
- Channel composition exploration
- Switch-patterns tailored for high-resistance lower metal
- MCNC benchmarks + VTR8.0

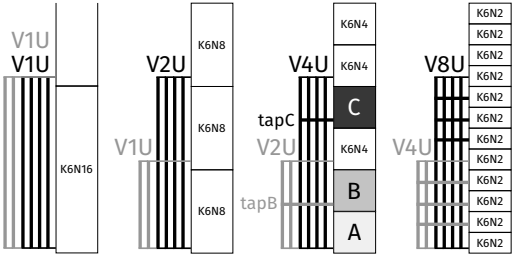


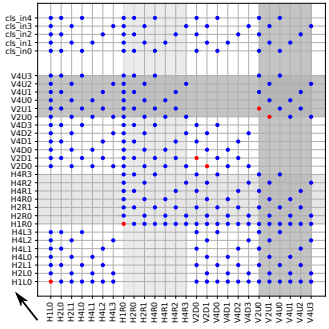


Experimental Setup

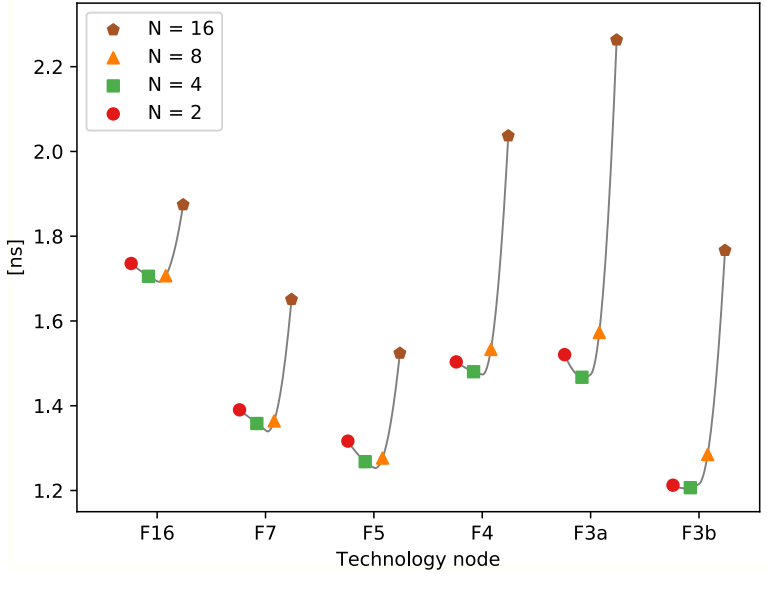
- Clusters of 2, 4, 8, and 16 6-LUTs
- Channel composition exploration
- Switch-patterns tailored for high-resistance lower metal
- MCNC benchmarks + VTR8.0

(Details in the paper)

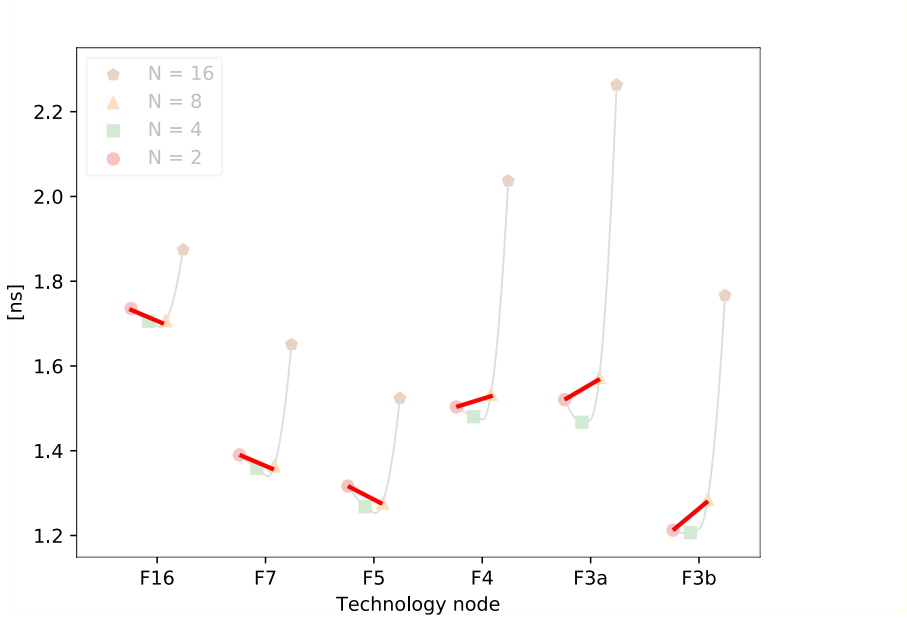




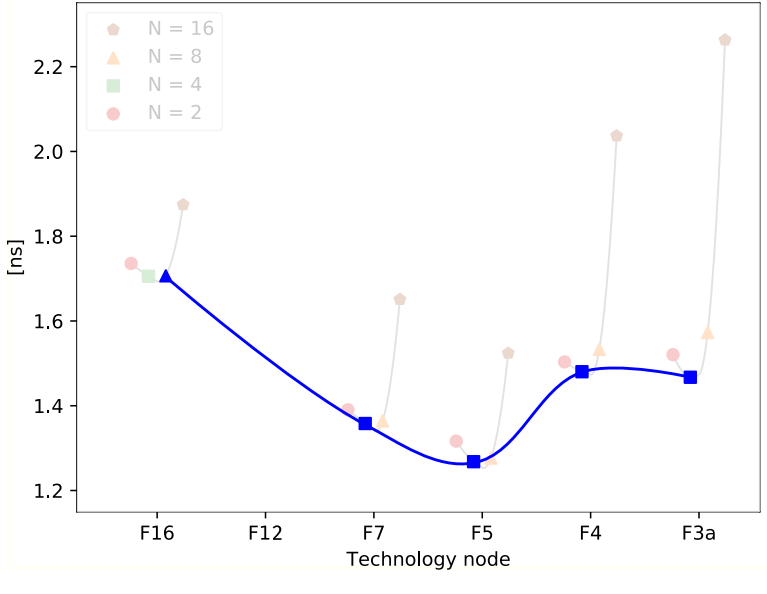
Cluster Sizes: Routed Delay Results



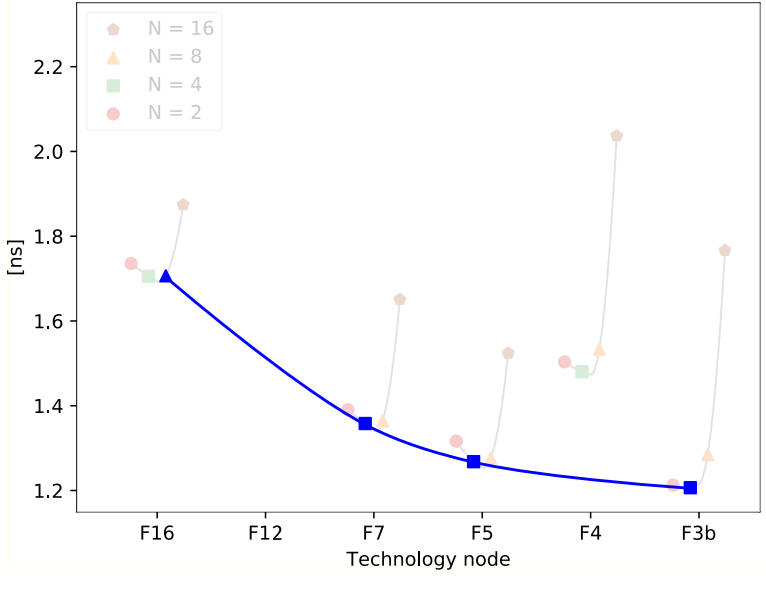
Cluster Sizes: Routed Delay Results



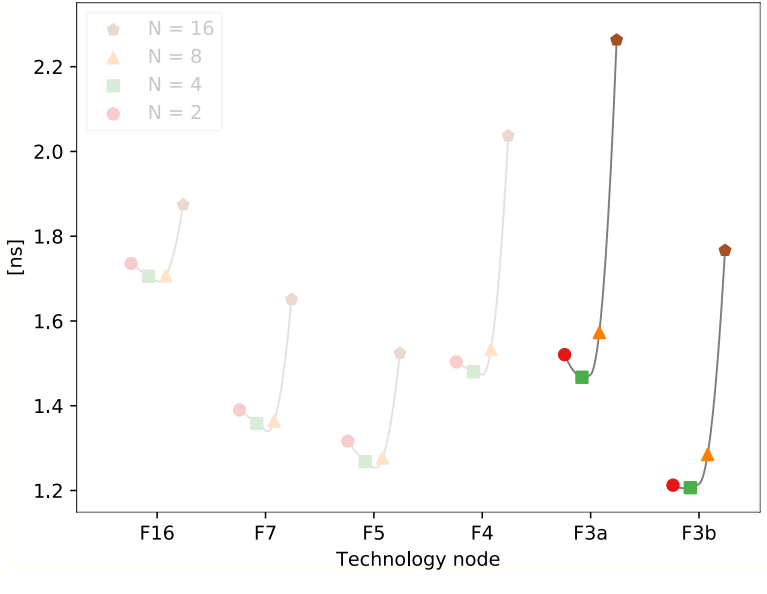
Performance Scaling



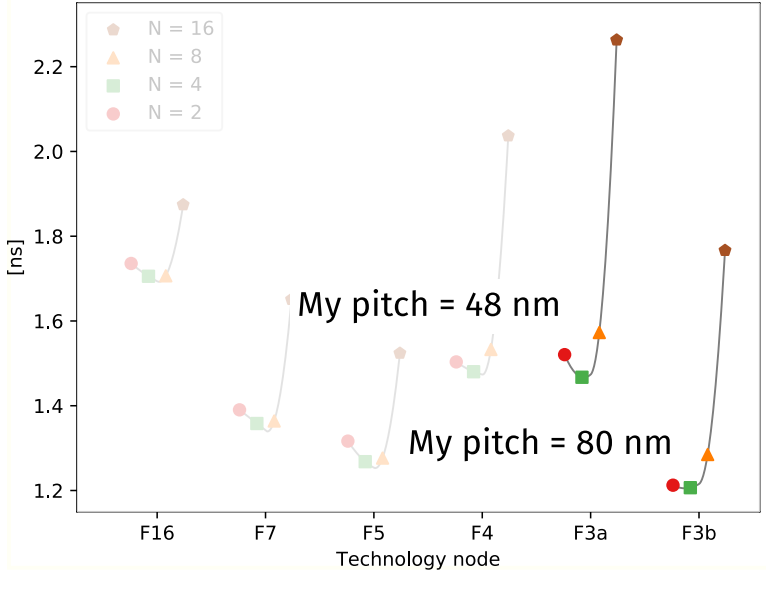
Performance Scaling



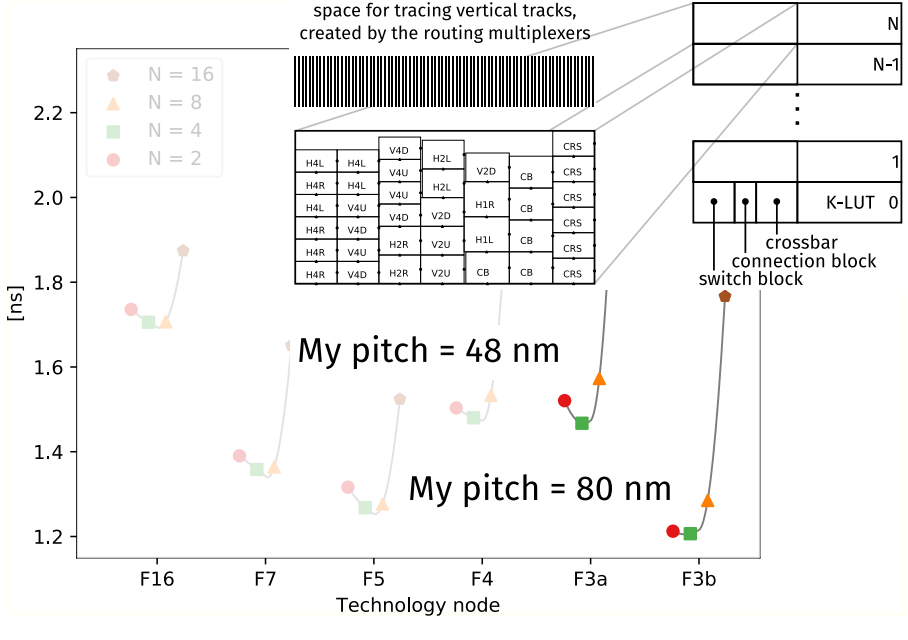
Performance Scaling



Performance Scaling



Performance Scaling



Performance Scaling

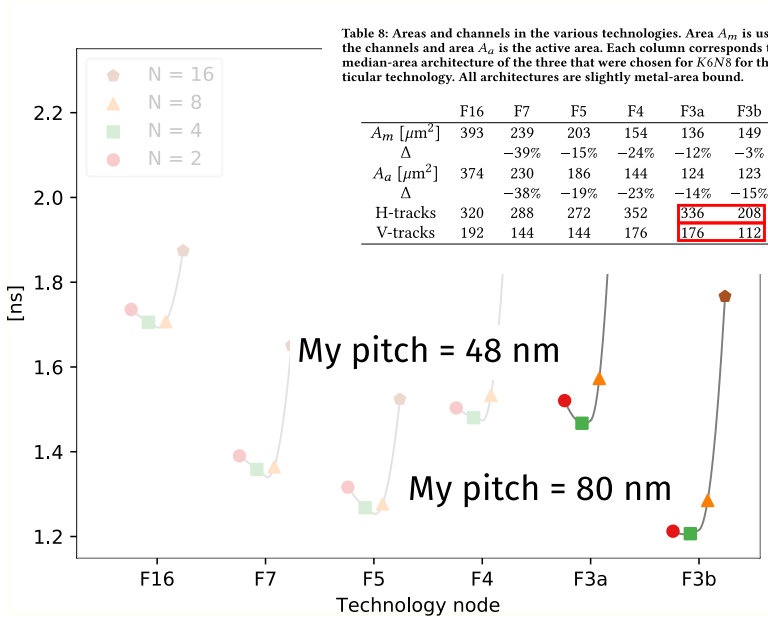


Table 8: Areas and channels in the various technologies. Area A_m is used by the channels and area A_a is the active area. Each column corresponds to the median-area architecture of the three that were chosen for $K6N8$ for the particular technology. All architectures are slightly metal-area bound.

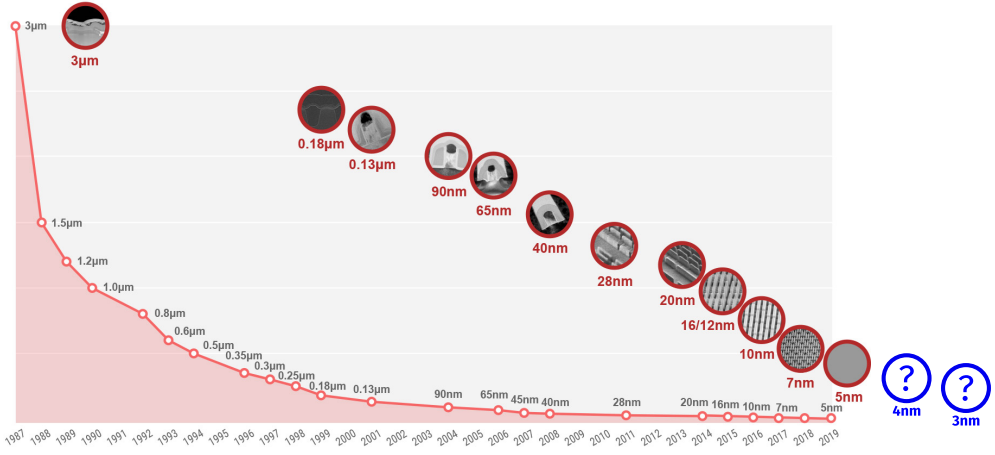
	F16	F7	F5	F4	F3a	F3b
A_m [μm^2]	393	239	203	154	136	149
Δ		−39%	−15%	−24%	−12%	−3%
A_a [μm^2]	374	230	186	144	124	123
Δ		−38%	−19%	−23%	−14%	−15%
H-tracks	320	288	272	352	336	208
V-tracks	192	144	144	176	176	112

-38%

-36%

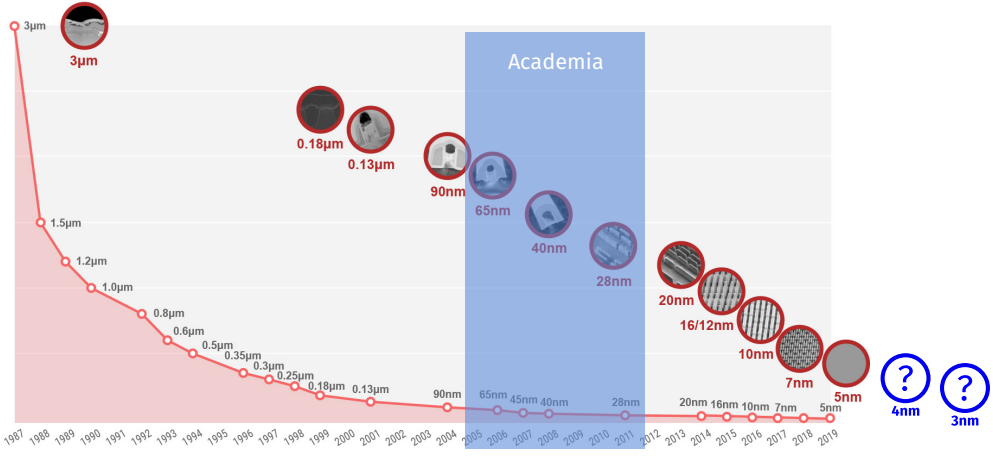
Summary

Summary



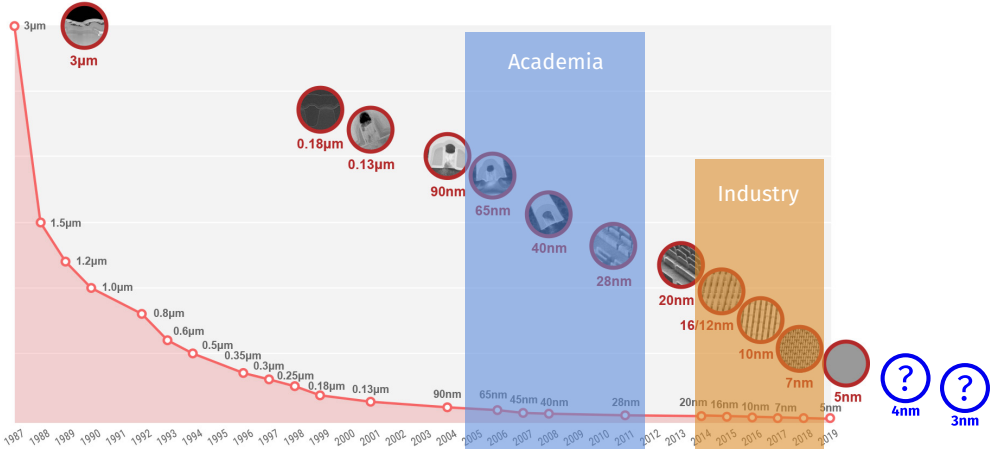
Source: TSMC (https://www.tsmc.com/english/dedicatedFoundry/technology/logic)

Summary



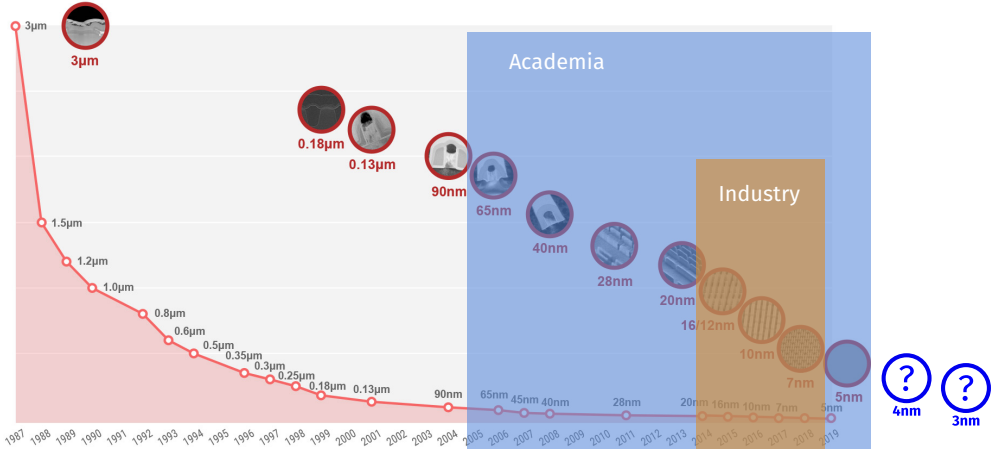
Source: TSMC (https://www.tsmc.com/english/dedicatedFoundry/technology/logic)

Summary



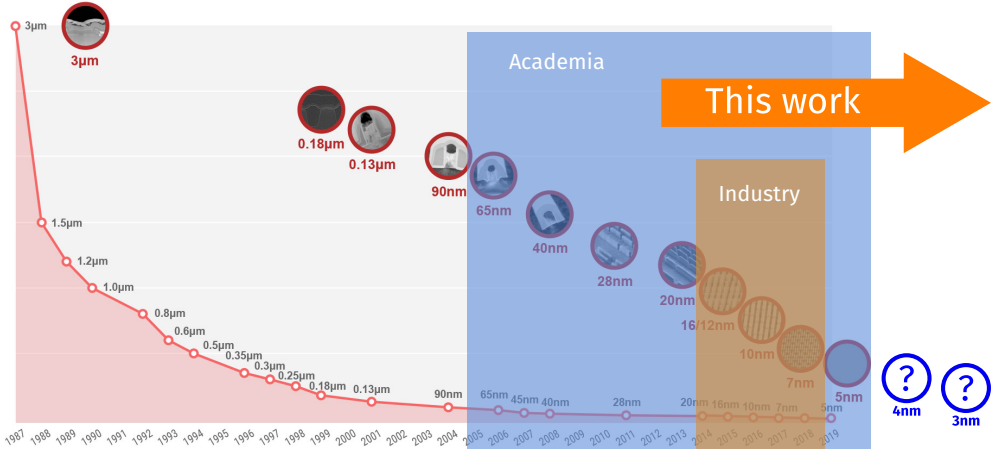
Source: TSMC (https://www.tsmc.com/english/dedicatedFoundry/technology/logic)

Summary



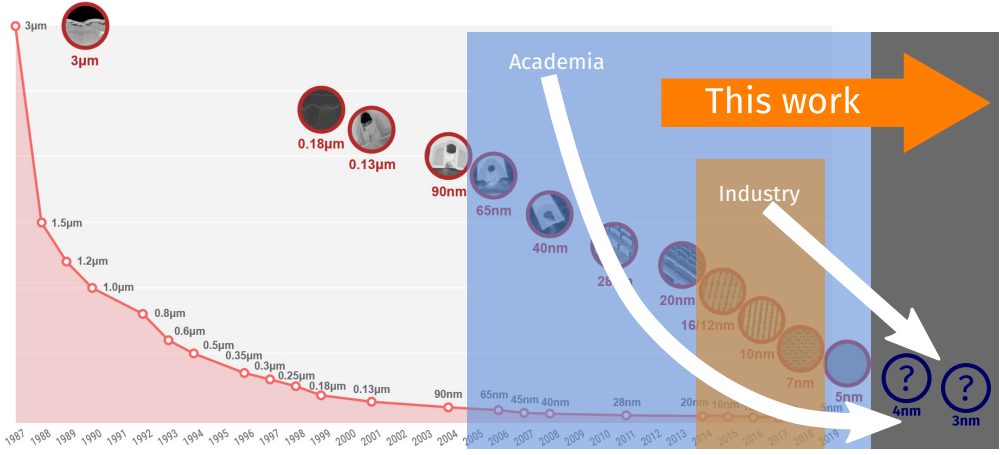
Source: TSMC (https://www.tsmc.com/english/dedicatedFoundry/technology/logic)

Summary



Source: TSMC (https://www.tsmc.com/english/dedicatedFoundry/technology/logic)

Summary



Source: TSMC (https://www.tsmc.com/english/dedicatedFoundry/technology/logic)

Thank you for attention

https://github.com/EPFL-LAP/fpga21-scaled-tech