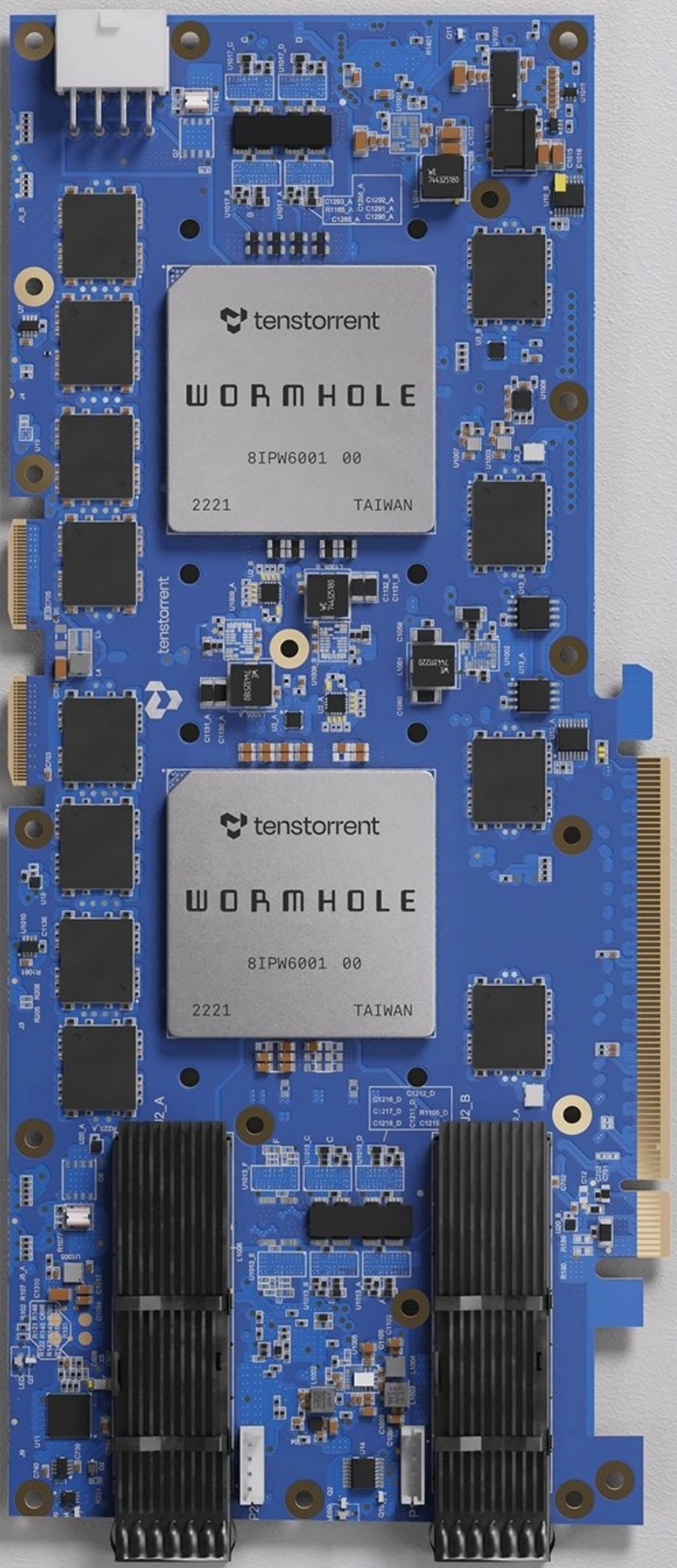
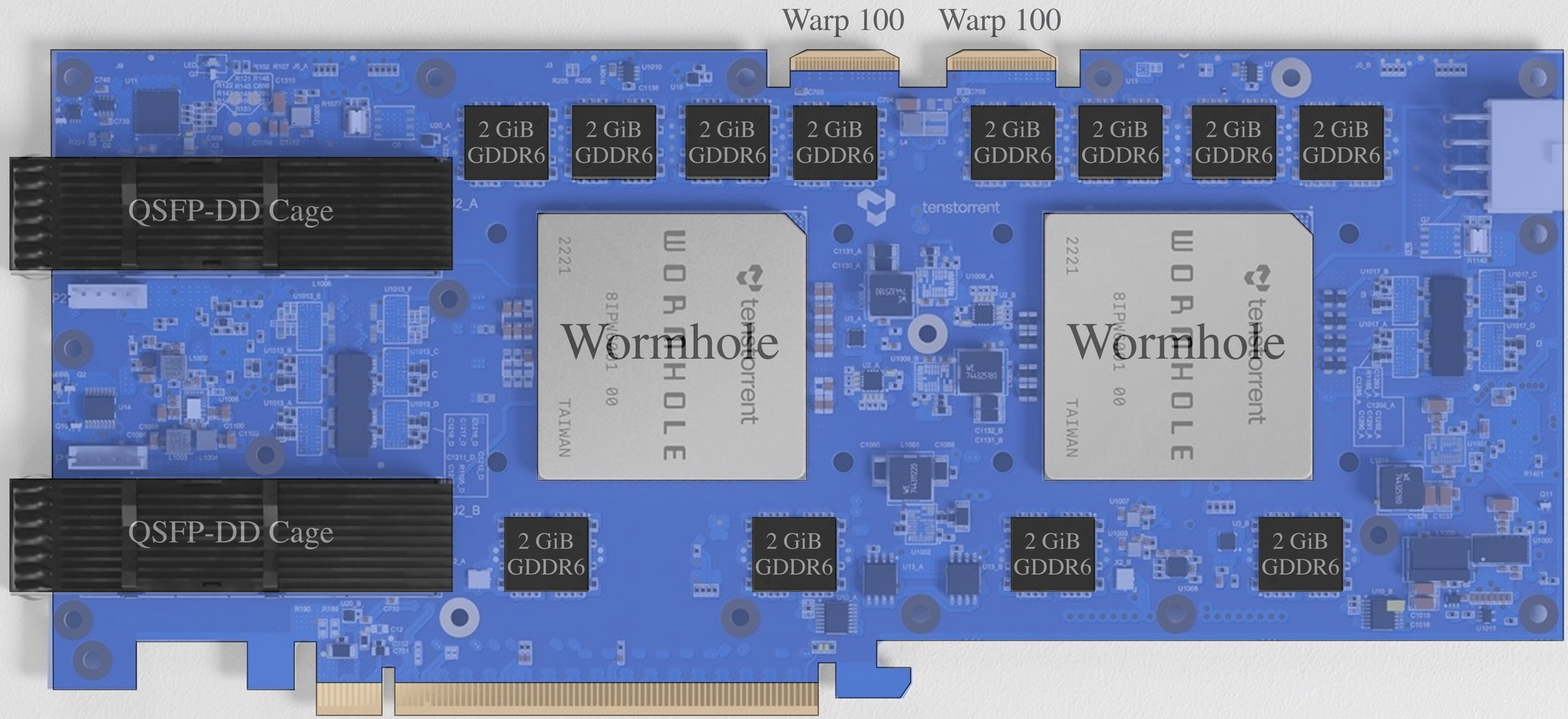
Scoping out the Tenstorrent Wormhole

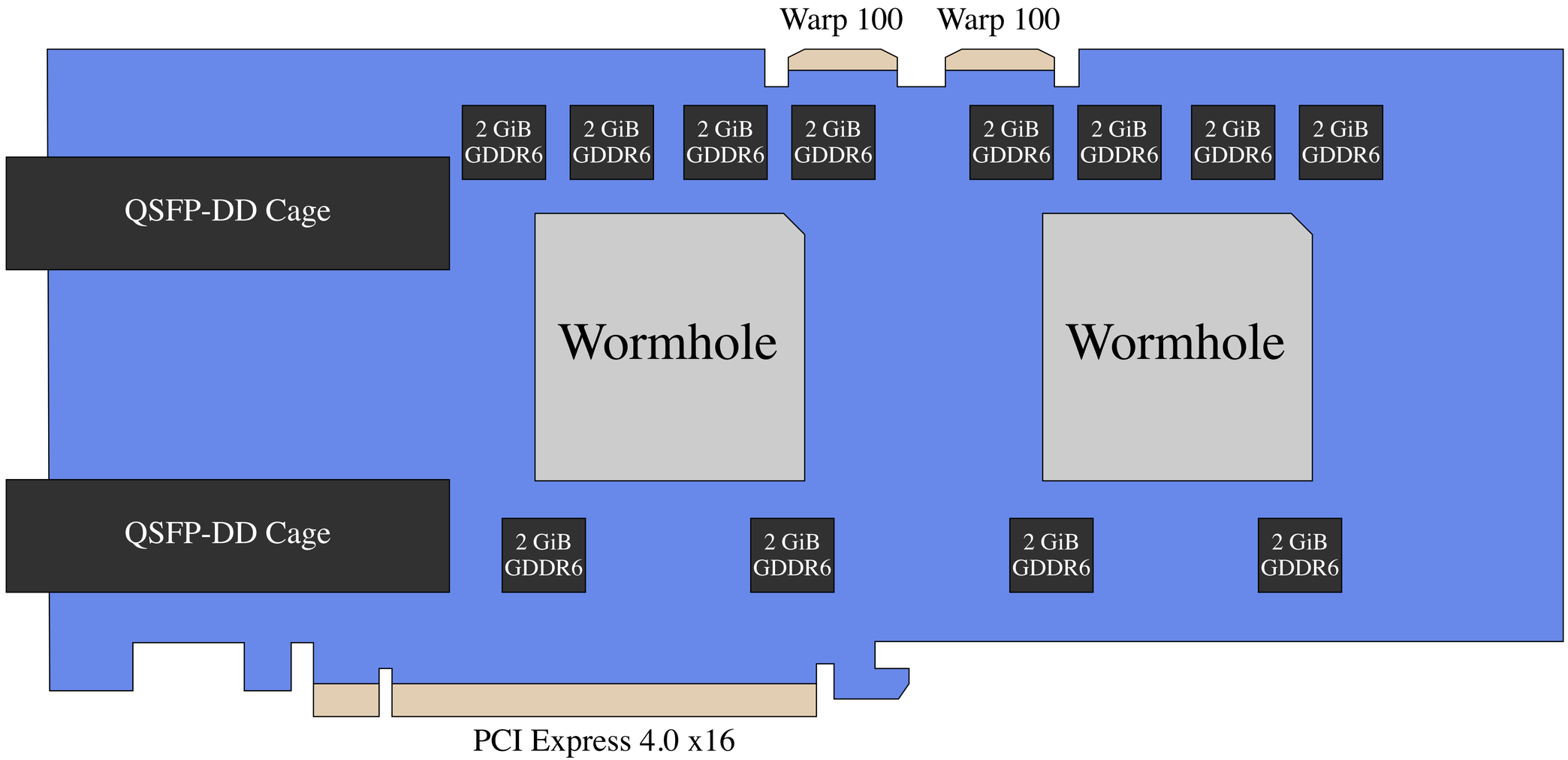
Peter Cawley (not affiliated with Tenstorrent … yet)

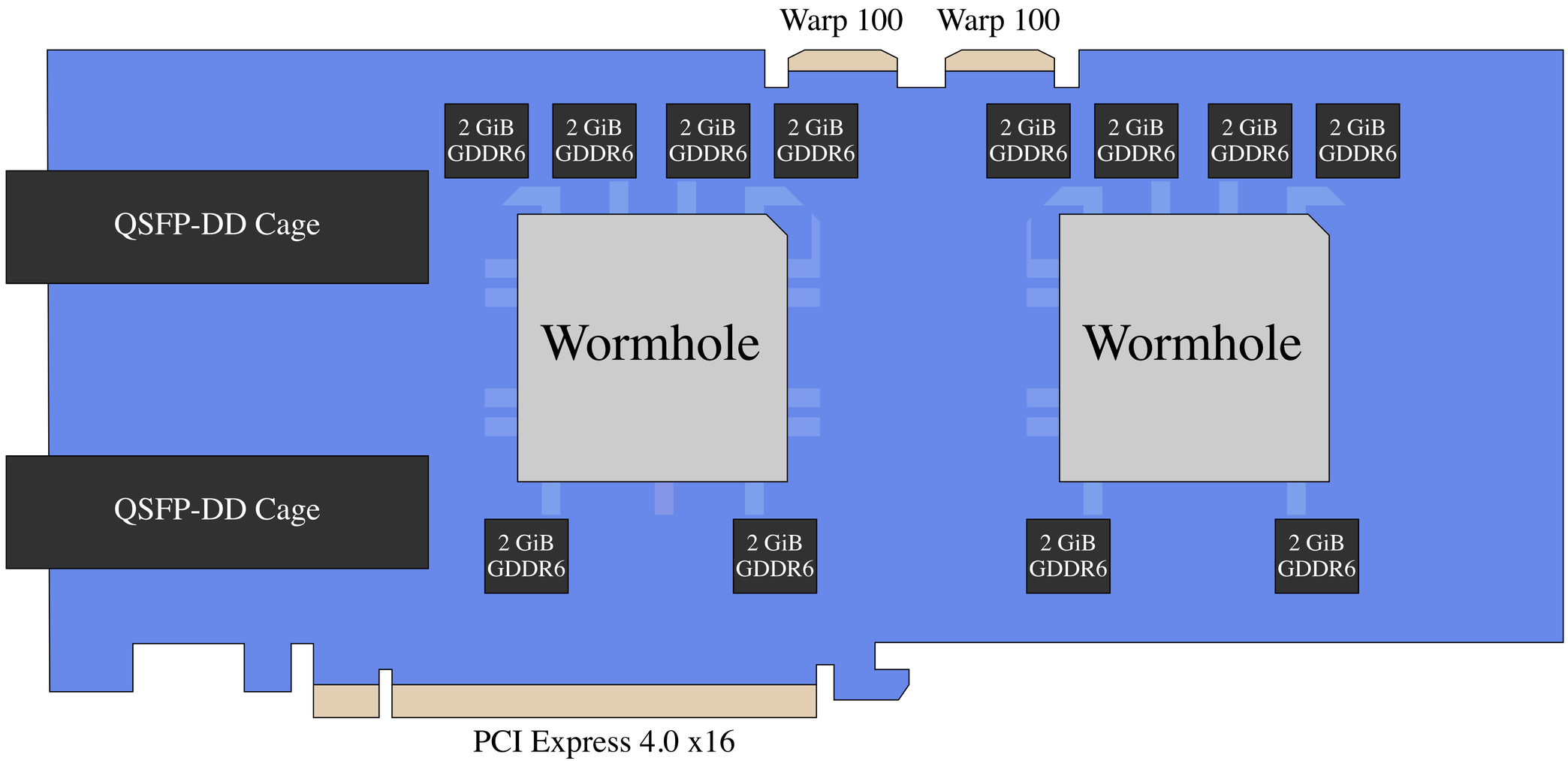
February 2nd, 2025 @ FOSDEM

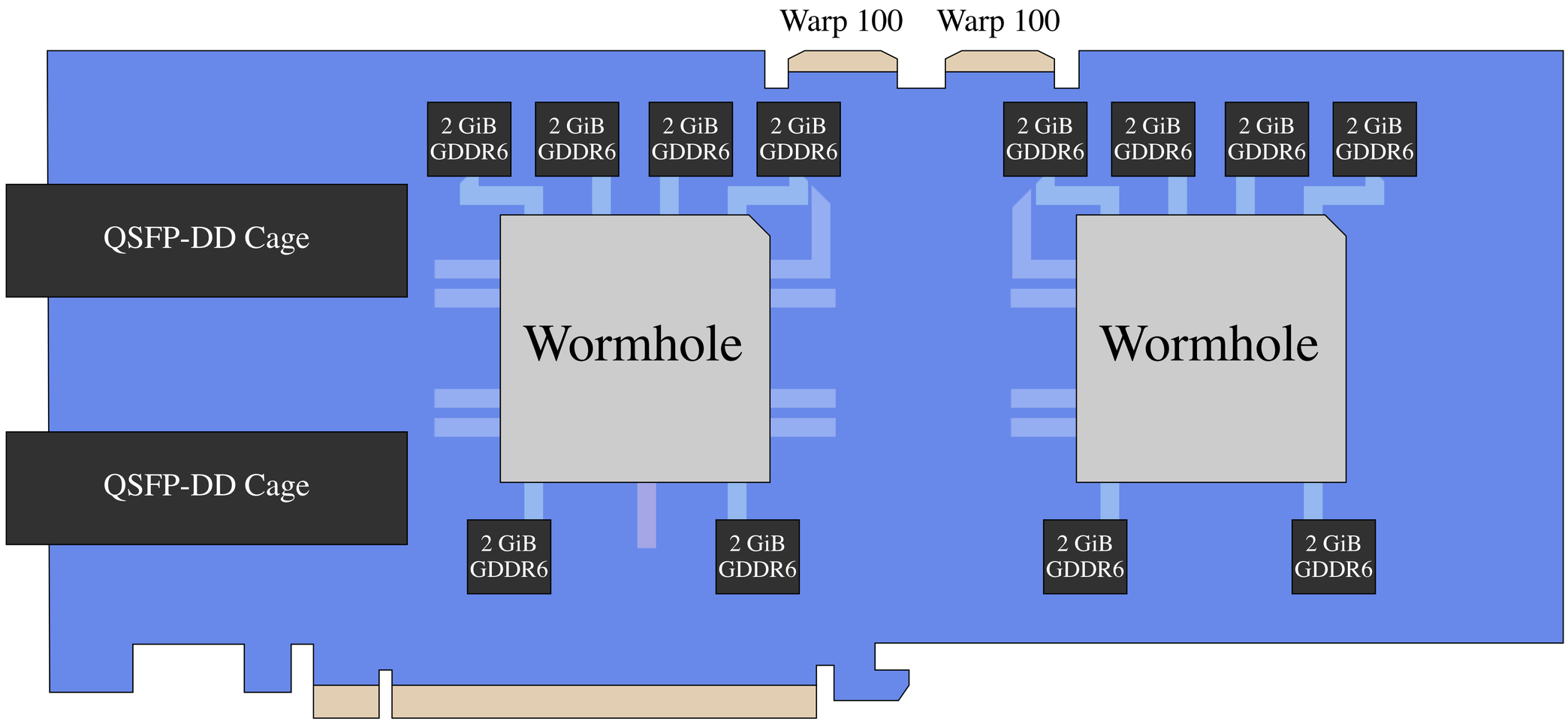
𝕏 @corsix

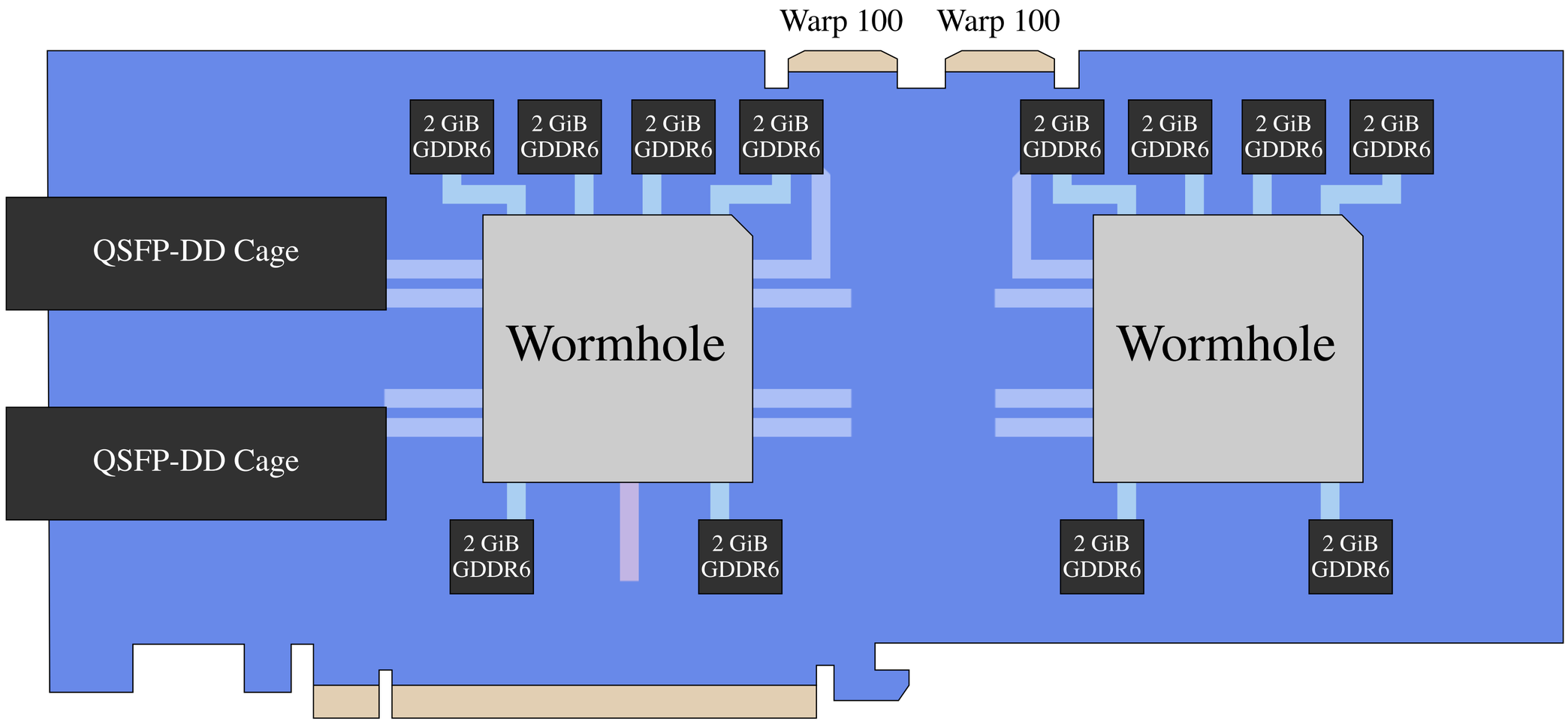

tenstorrent
WORMHOLE
8IPW6001 00
2221 TAIWAN
tenstorrent
WORMHOLE
8IPW6001 00
2221 TAIWAN

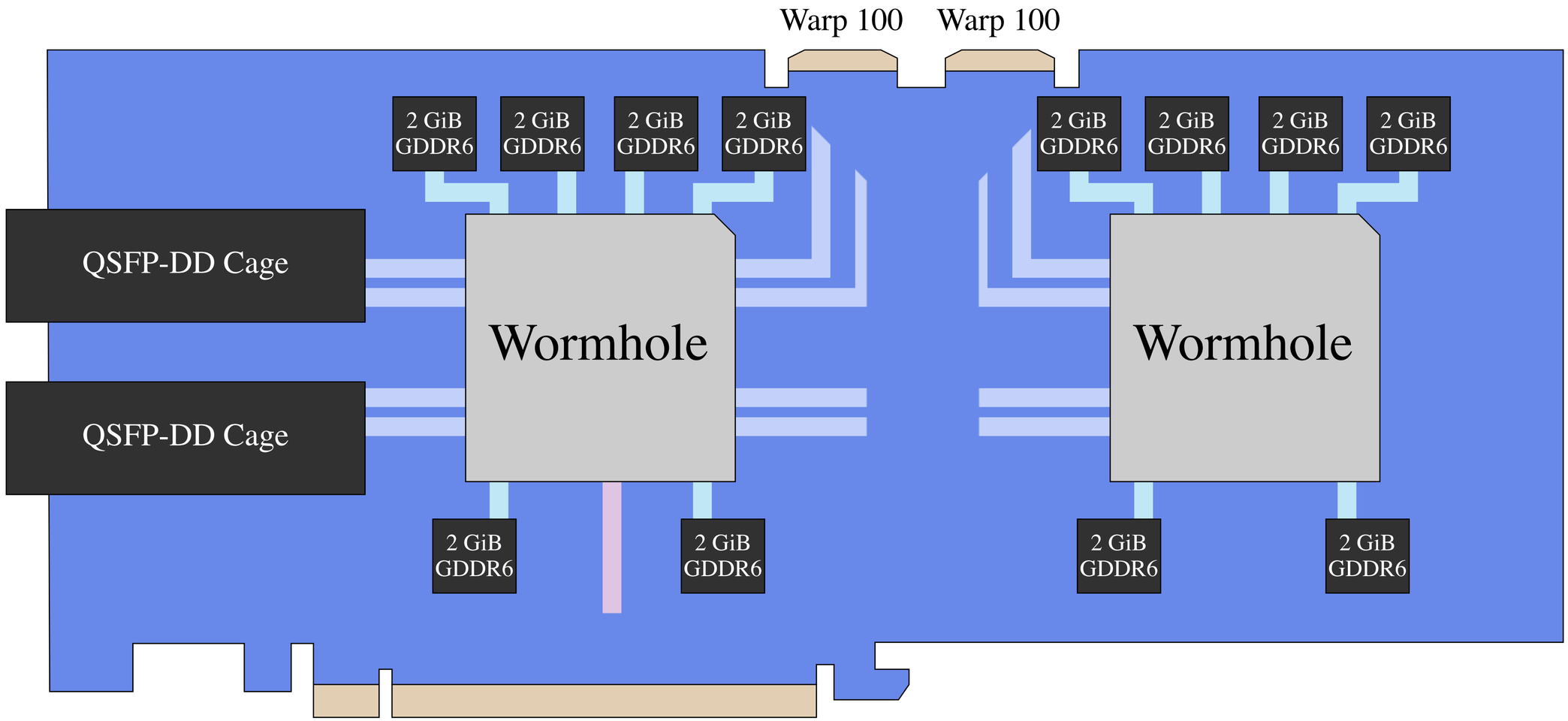

Warp 100
Warp 100
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
QSFP-DD Cage
Wormhole
Wormhole
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
QSFP-DD Cage
PCI Express 4.0 x16

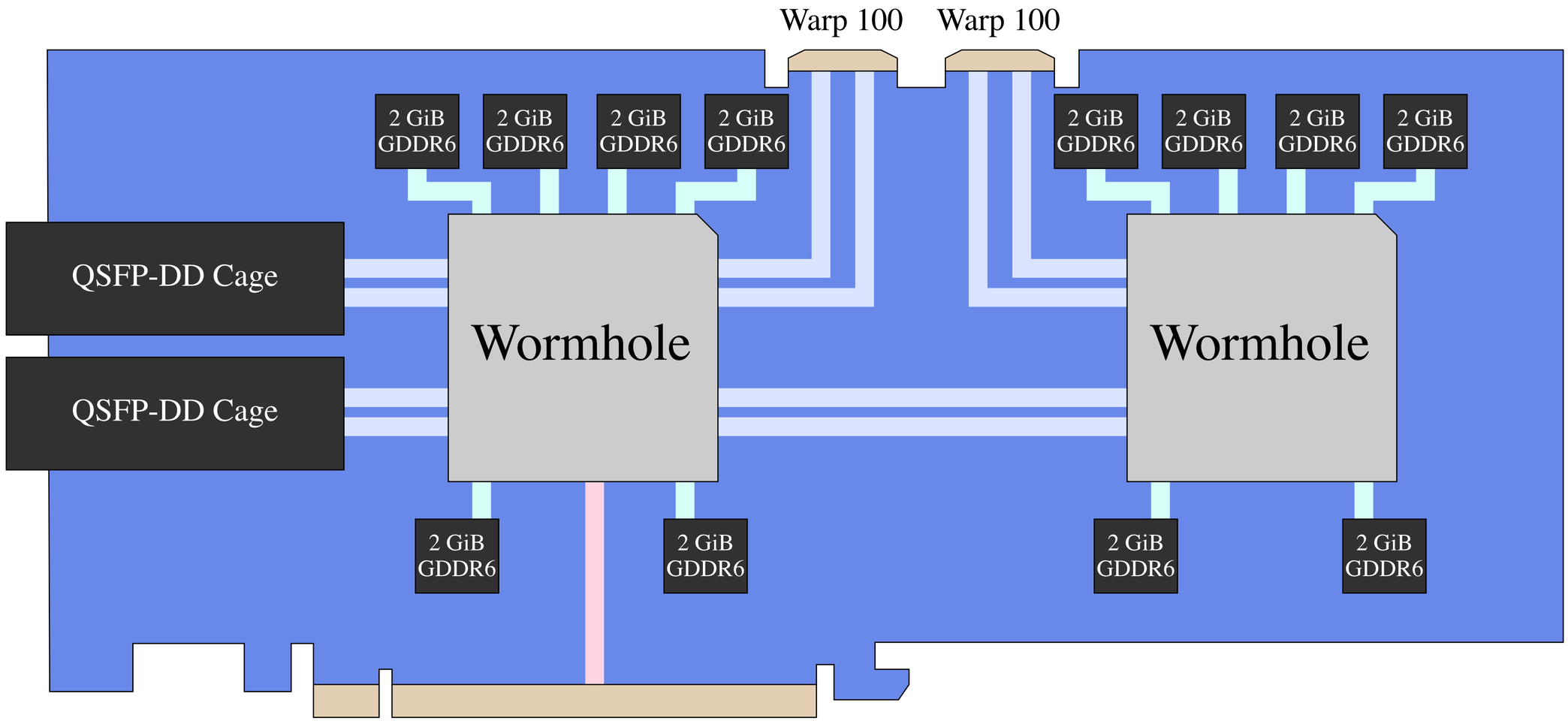

Warp 100
Warp 100
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
QSFP-DD Cage
Wormhole
Wormhole
QSFP-DD Cage
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
PCI Express 4.0 x16

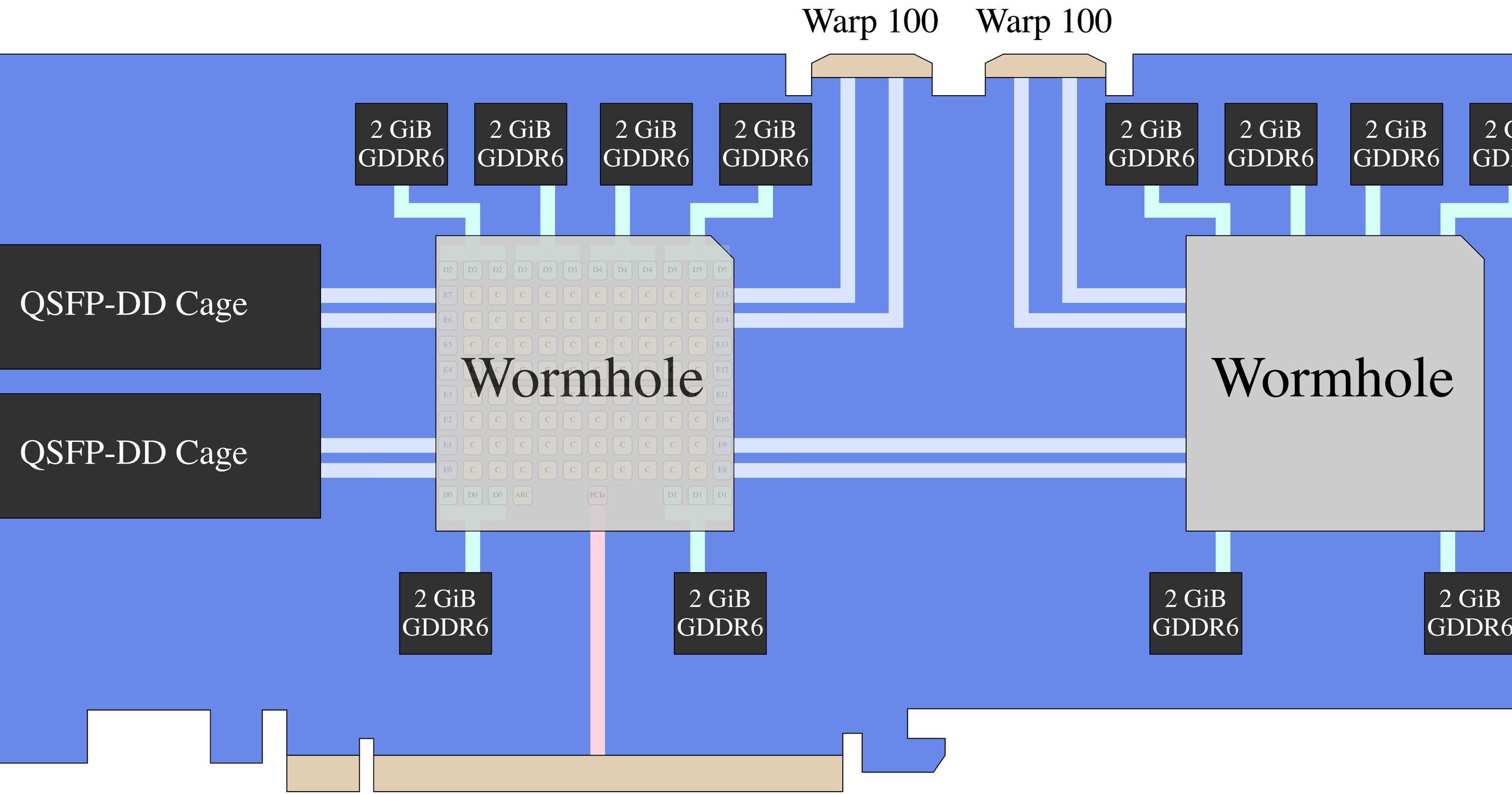

Warp 100
Warp 100
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
QSFP-DD Cage
Wormhole
Wormhole
QSFP-DD Cage
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
PCI Express 4.0 x16


Warp 100
Warp 100
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
QSFP-DD Cage
Wormhole
Wormhole
QSFP-DD Cage
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
PCI Express 4.0 x16


Warp 100
Warp 100
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
QSFP-DD Cage
QSFP-DD Cage
Wormhole
Wormhole
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
PCI Express 4.0 x16


Warp 100
Warp 100
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
QSFP-DD Cage
QSFP-DD Cage
Wormhole
Wormhole
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
PCI Express 4.0 x16


Warp 100
Warp 100
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
QSFP-DD Cage
QSFP-DD Cage
Wormhole
Wormhole
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
PCI Express 4.0 x16


Warp 100
Warp 100
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
QSFP-DD Cage
QSFP-DD Cage
Wormhole
Wormhole
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
PCI Express 4.0 x16

Warp 100 Warp 100

2 GiB GDDR6

2 GiB GDDR6

2 GiB GDDR6

2 GiB GDDR6

2 GiB GDDR6

2 GiB GDDR6

DD Cage

DD Cage

Wormhole

Wor

2 GiB GDDR6

2 GiB GDDR6

2 GiB GDDR6

3.4

2 GiB GDDR6

2 GiB GDDR6

2 GiB GDDR6

2 GiB GDDR6

ge

ge

D2	D2	D2	D3	D3	D3	D4	D4	D4	D5	D5	D5
E7	C	C	C	C	C	C	C	C	C	C	E15
E6	C	C	C	C	C	C	C	C	C	C	E14
E5	C	C	C	C	C	C	C	C	C	C	E13
E4	C	C	C	C	C	C	C	C	C	C	E12
E3	C	C	C	C	C	C	C	C	C	C	E11
E2	C	C	C	C	C	C	C	C	C	C	E10
E1	C	C	C	C	C	C	C	C	C	C	E9
E0	C	C	C	C	C	C	C	C	C	C	E8
D0	D0	D0	ARC			PCIe			D1	D1	D1

Wormhole

2 GiB GDDR6

2 GiB GDDR6

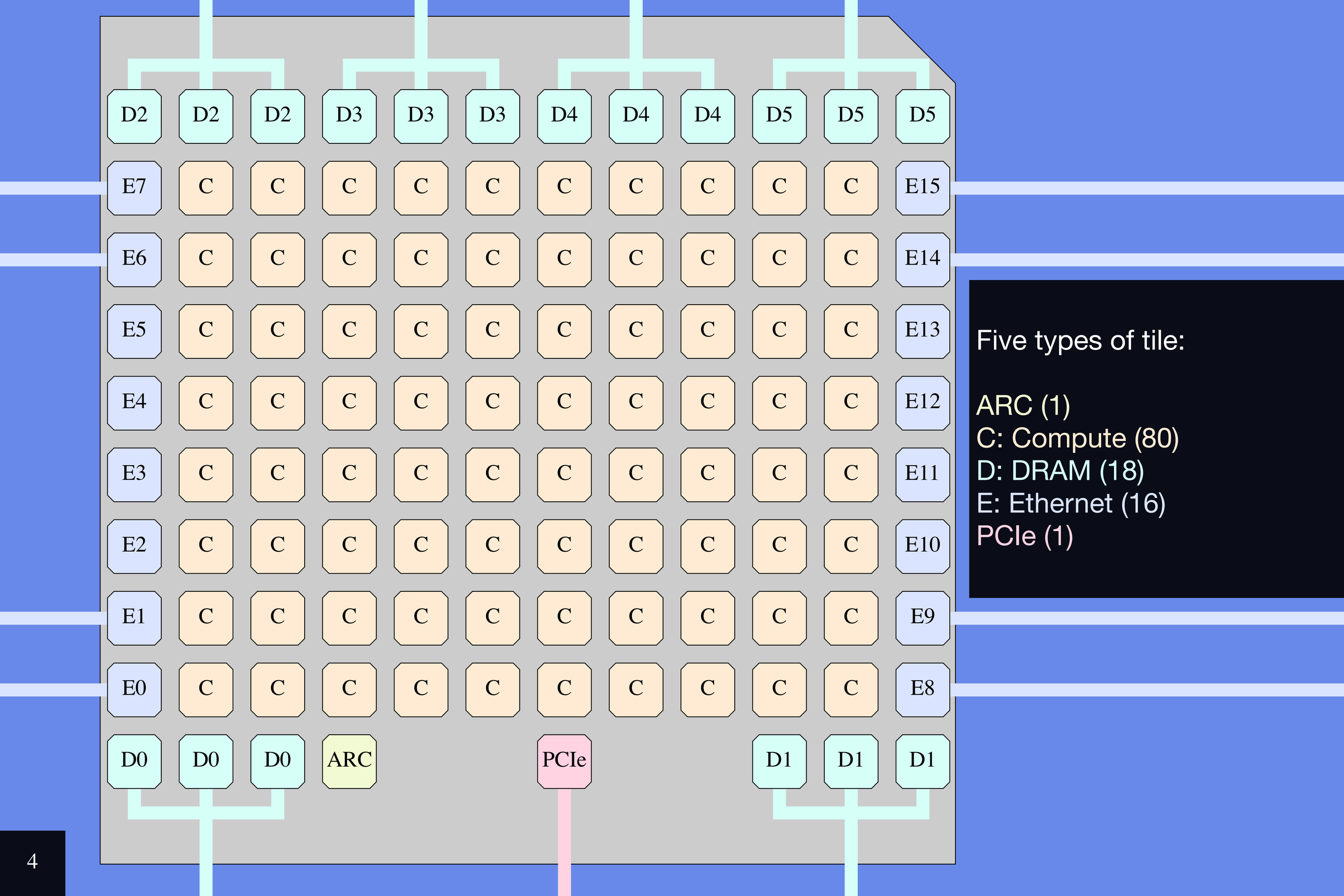
3.6

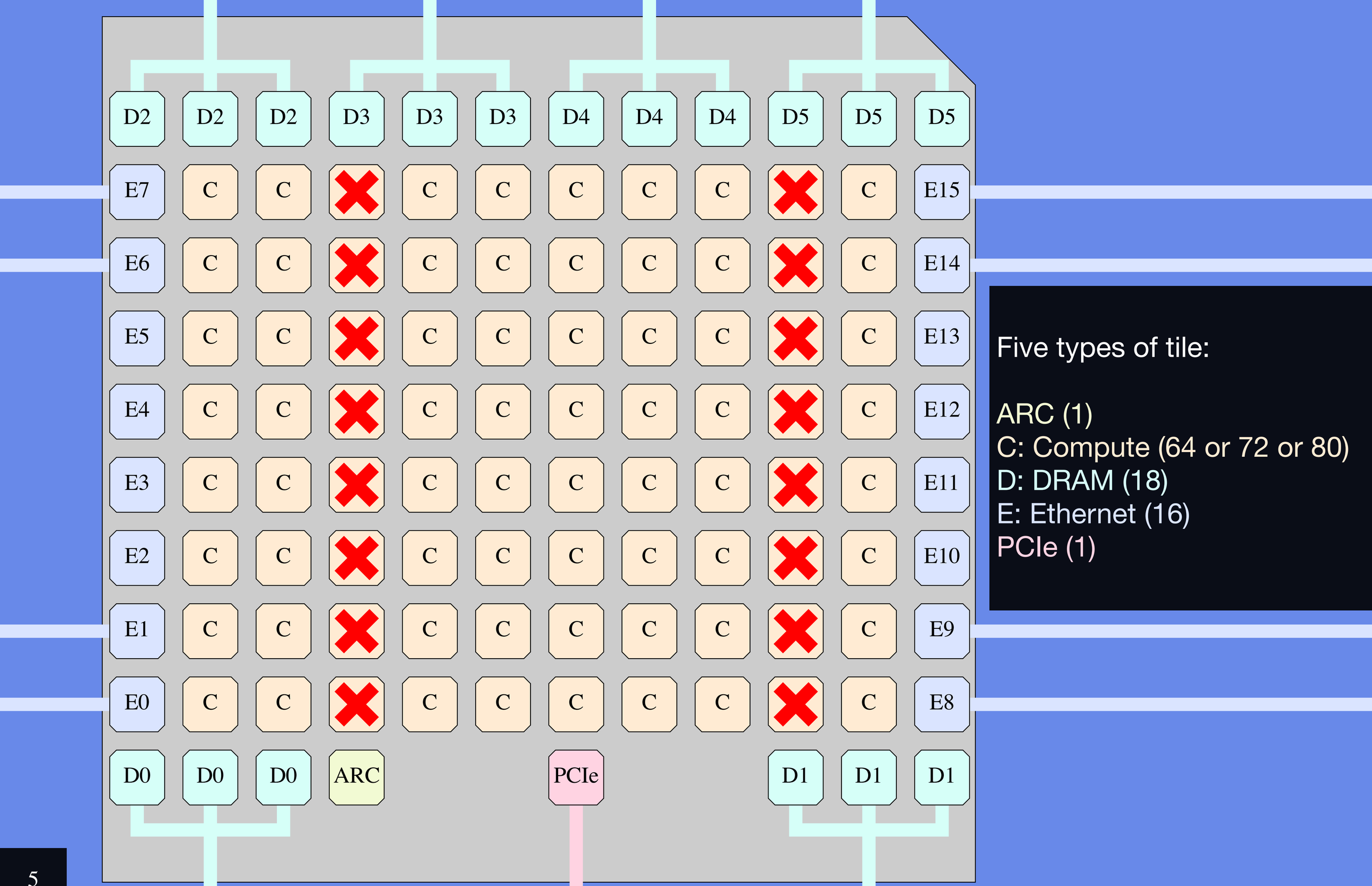
GDDR6 GDDR6 GDDR6 GDDR6

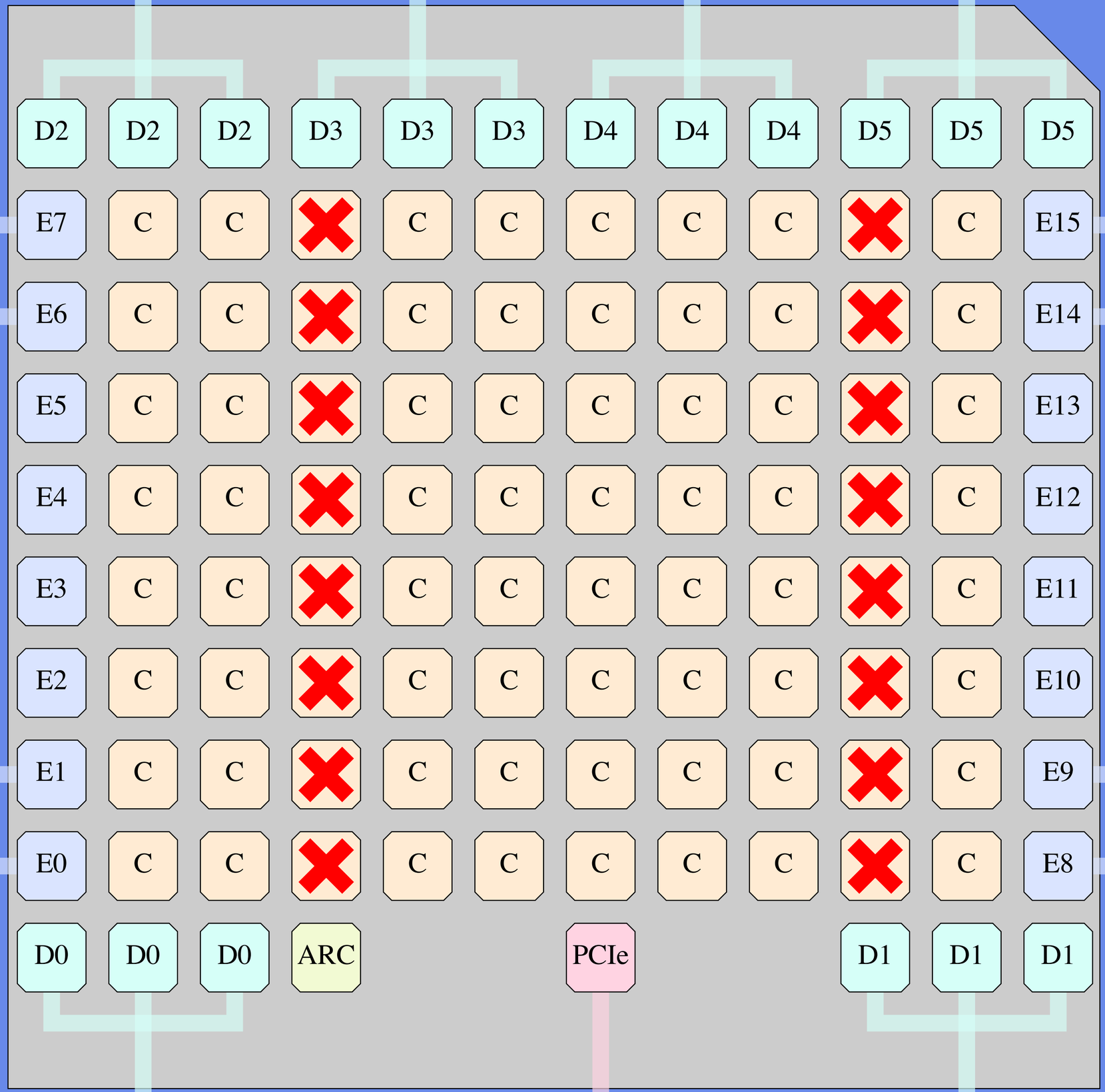
D2	D2	D2	D3	D3	D3	D4	D4	D4	D5	D5	D5
E7	C	C	C	C	C	C	C	C	C	C	E15
E6	C	C	C	C	C	C	C	C	C	C	E14
E5	C	C	C	C	C	C	C	C	C	C	E13
E4	C	C	C	C	C	C	C	C	C	C	E12
E3	C	C	C	C	C	C	C	C	C	C	E11
E2	C	C	C	C	C	C	C	C	C	C	E10
E1	C	C	C	C	C	C	C	C	C	C	E9
E0	C	C	C	C	C	C	C	C	C	C	E8
D0	D0	D0	ARC		PCIe				D1	D1	D1

2 GiB

2 GiB

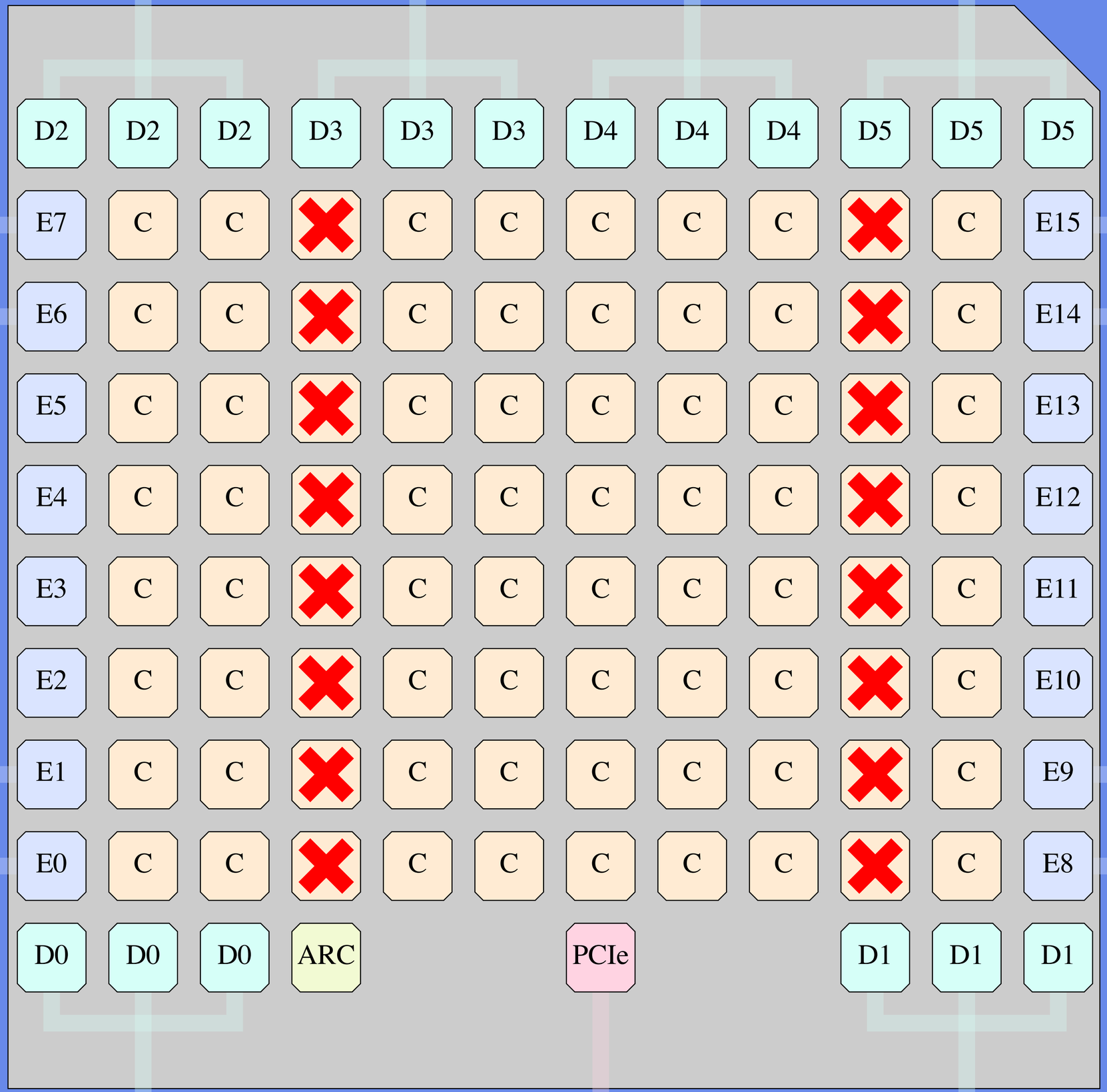

D2 D2 D2 D3 D3 D3 D4 D4 D4 D5 D5 D5
E7 C C C C C C C C C C E15
E6 C C C C C C C C C C E14
E5 C C C C C C C C C C E13
E4 C C C C C C C C C C E12
E3 C C C C C C C C C C E11
E2 C C C C C C C C C C E10
E1 C C C C C C C C C C E9
E0 C C C C C C C C C C E8
D0 D0 D0 ARC PCIe D1 D1 D1
Five types of tile:
ARC (1)
C: Compute (80)
D: DRAM (18)
E: Ethernet (16)
PCIe (1)


D2 D2 D2 D3 D3 D3 D4 D4 D4 D5 D5 D5
E7 C C C C C C C C C E15
E6 C C C C C C C C C E14
E5 C C C C C C C C C E13
E4 C C C C C C C C C E12
E3 C C C C C C C C C E11
E2 C C C C C C C C C E10
E1 C C C C C C C C C E9
E0 C C C C C C C C C E8
D0 D0 D0 ARC PCIe D1 D1 D1
Five types of tile:
ARC (1)
C: Compute (64 or 72 or 80)
D: DRAM (18)
E: Ethernet (16)
PCIe (1)



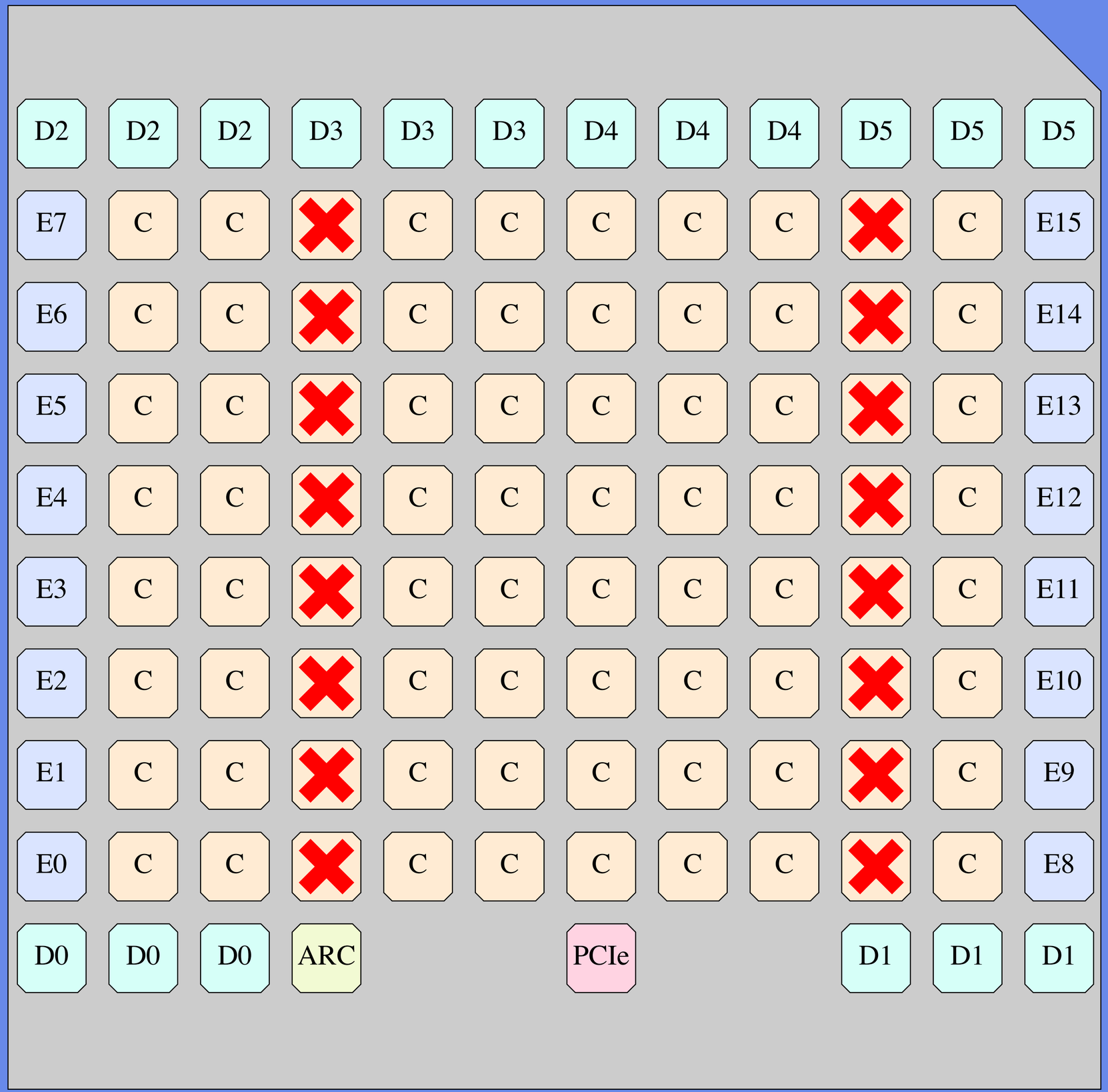
Five types of tile:

ARC (1)
C: Compute (64 or 72 or 80)
D: DRAM (18)
E: Ethernet (16)
PCIe (1)



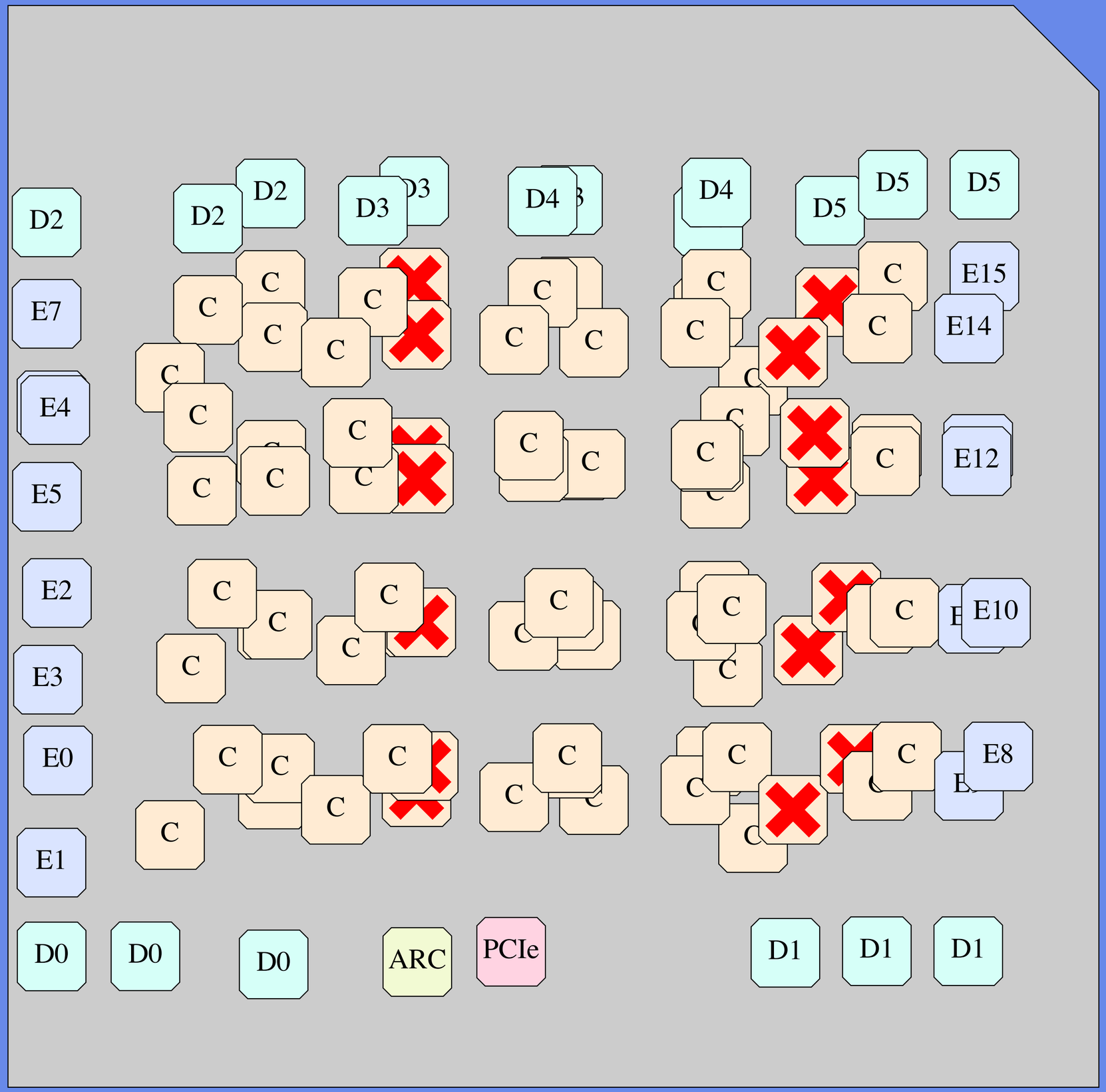
Five types of tile:

ARC (1)
C: Compute (64 or 72 or 80)
D: DRAM (18)
E: Ethernet (16)
PCIe (1)



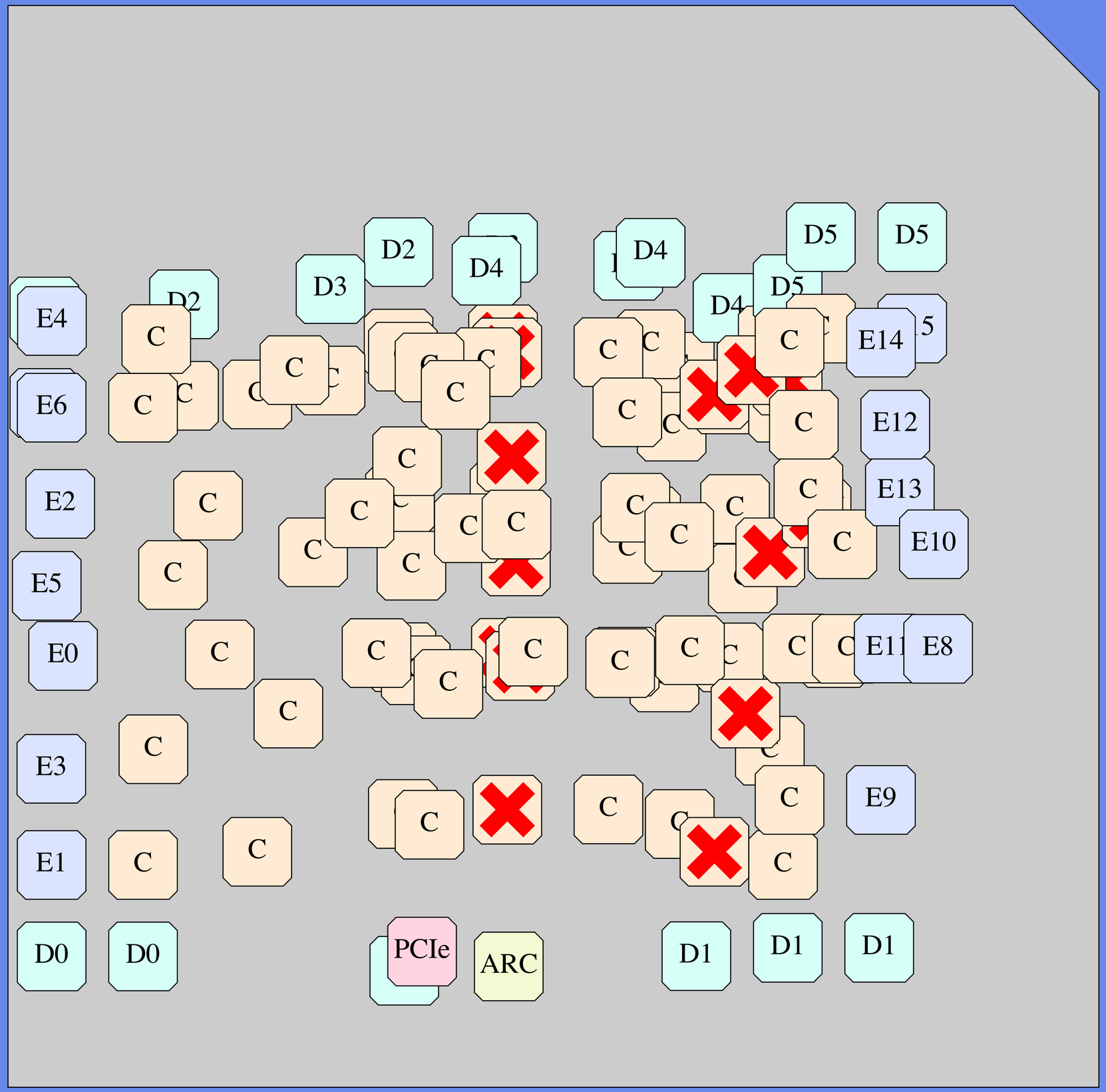
Five types of tile:

ARC (1)
C: Compute (64 or 72 or 80)
D: DRAM (18)
E: Ethernet (16)
PCIe (1)



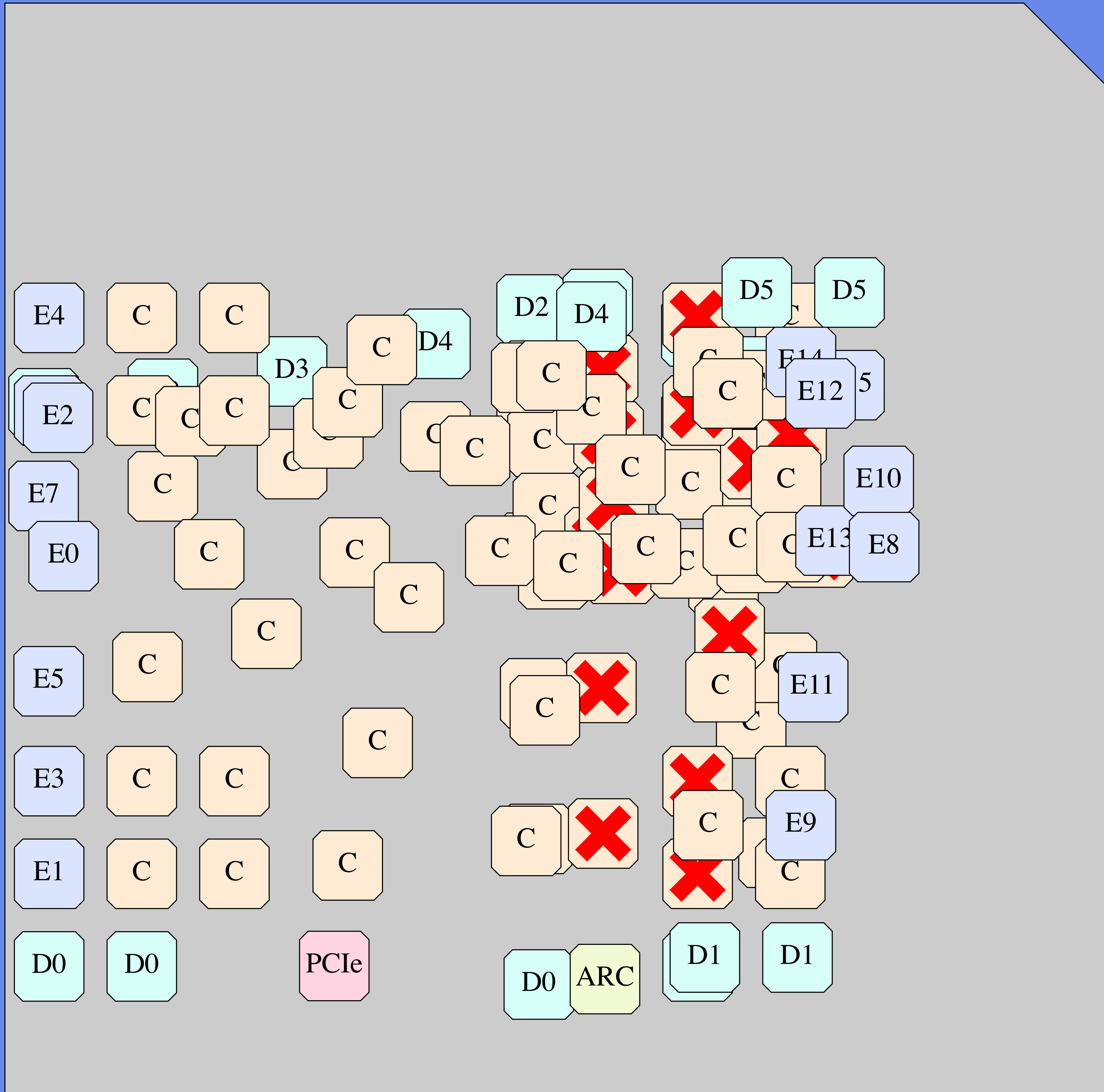
Five types of tile:

ARC (1)
C: Compute (64 or 72 or 80)
D: DRAM (18)
E: Ethernet (16)
PCIe (1)



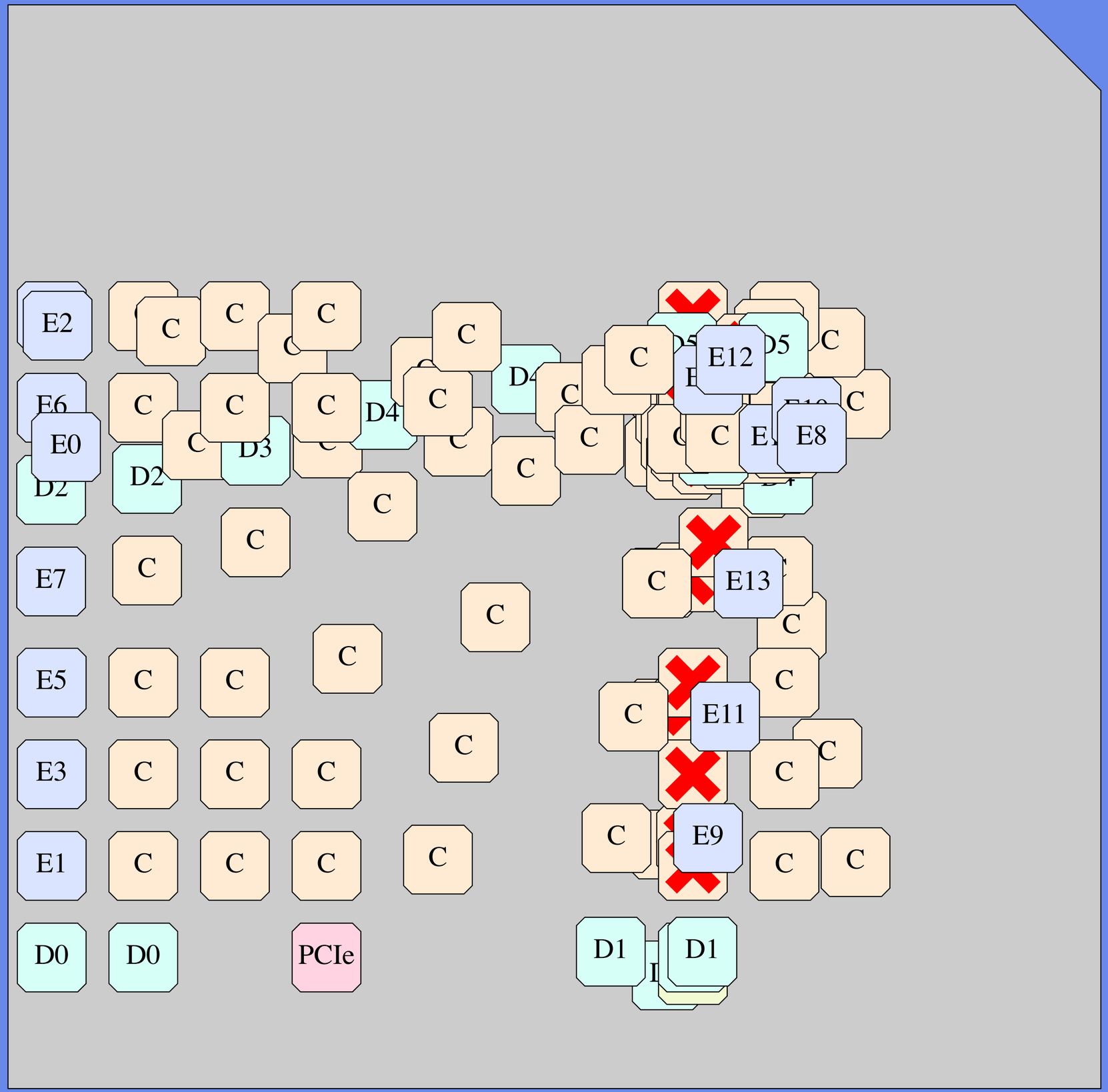
Five types of tile:

ARC (1)
C: Compute (64 or 72 or 80)
D: DRAM (18)
E: Ethernet (16)
PCIe (1)



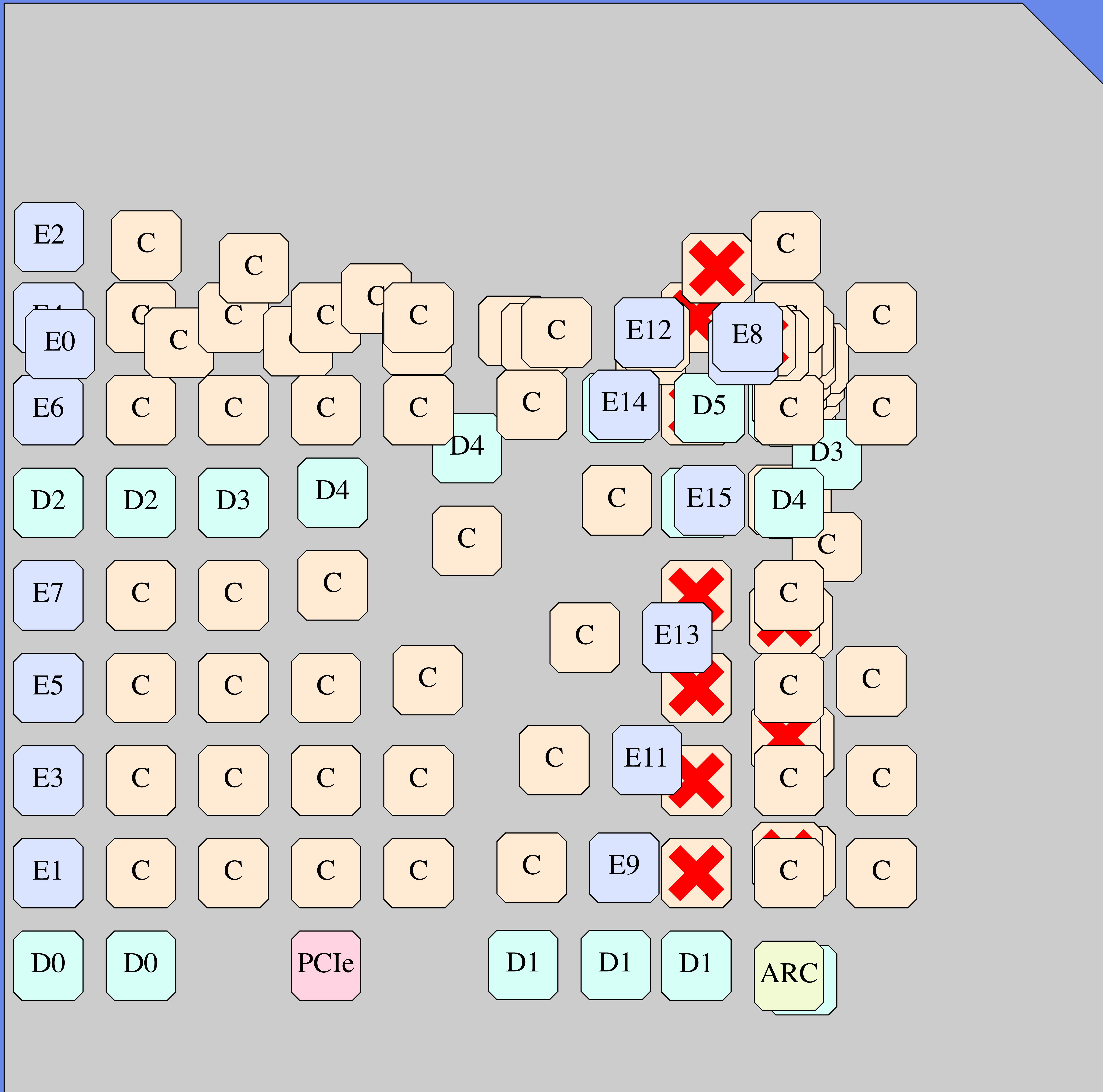
Five types of tile:

ARC (1)
C: Compute (64 or 72 or 80)
D: DRAM (18)
E: Ethernet (16)
PCIe (1)



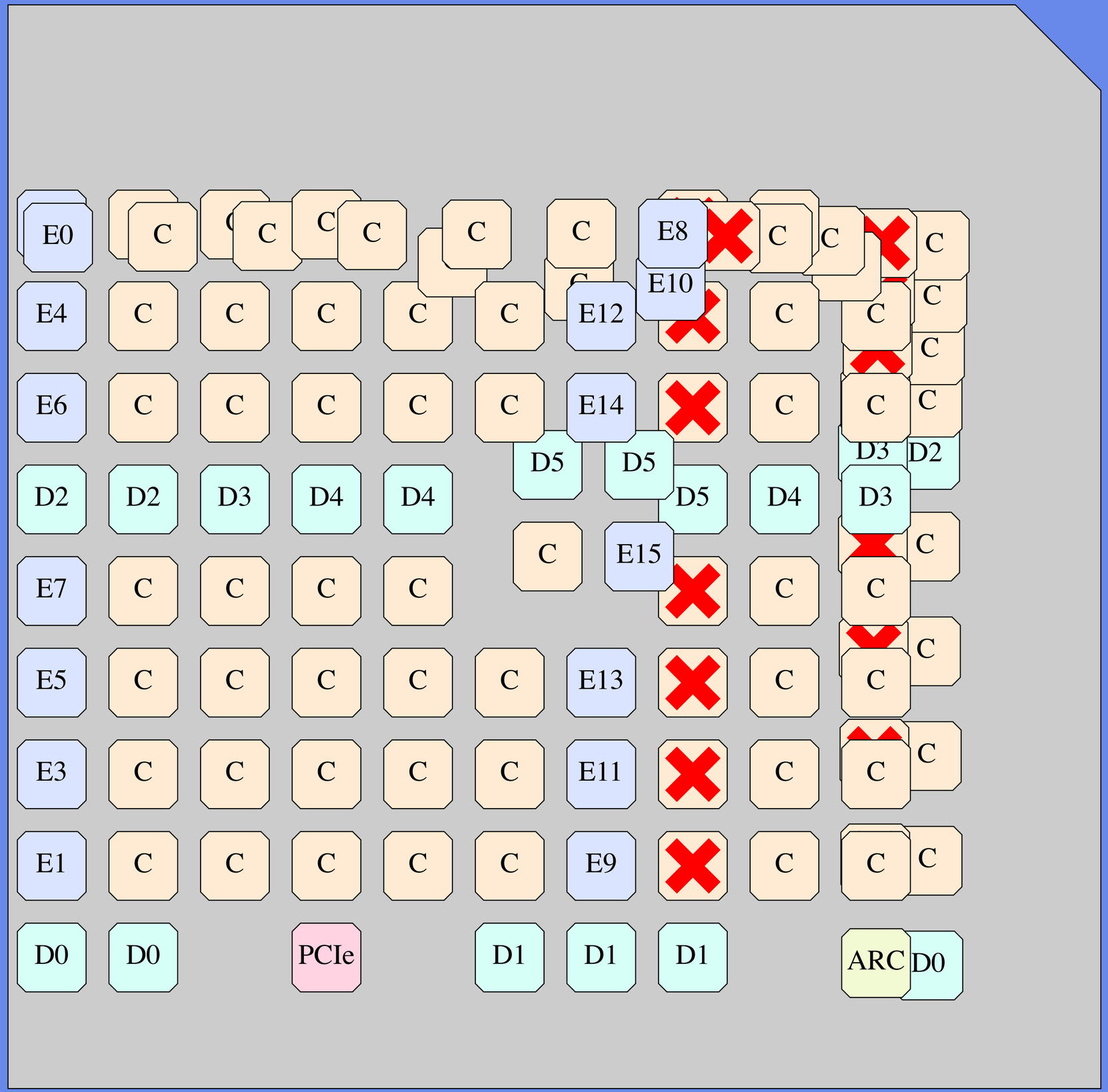
Five types of tile:

ARC (1)
C: Compute (64 or 72 or 80)
D: DRAM (18)
E: Ethernet (16)
PCIe (1)



Five types of tile:

ARC (1)
C: Compute (64 or 72 or 80)
D: DRAM (18)
E: Ethernet (16)
PCIe (1)



Five types of tile:

ARC (1)
C: Compute (64 or 72 or 80)
D: DRAM (18)
E: Ethernet (16)
PCIe (1)

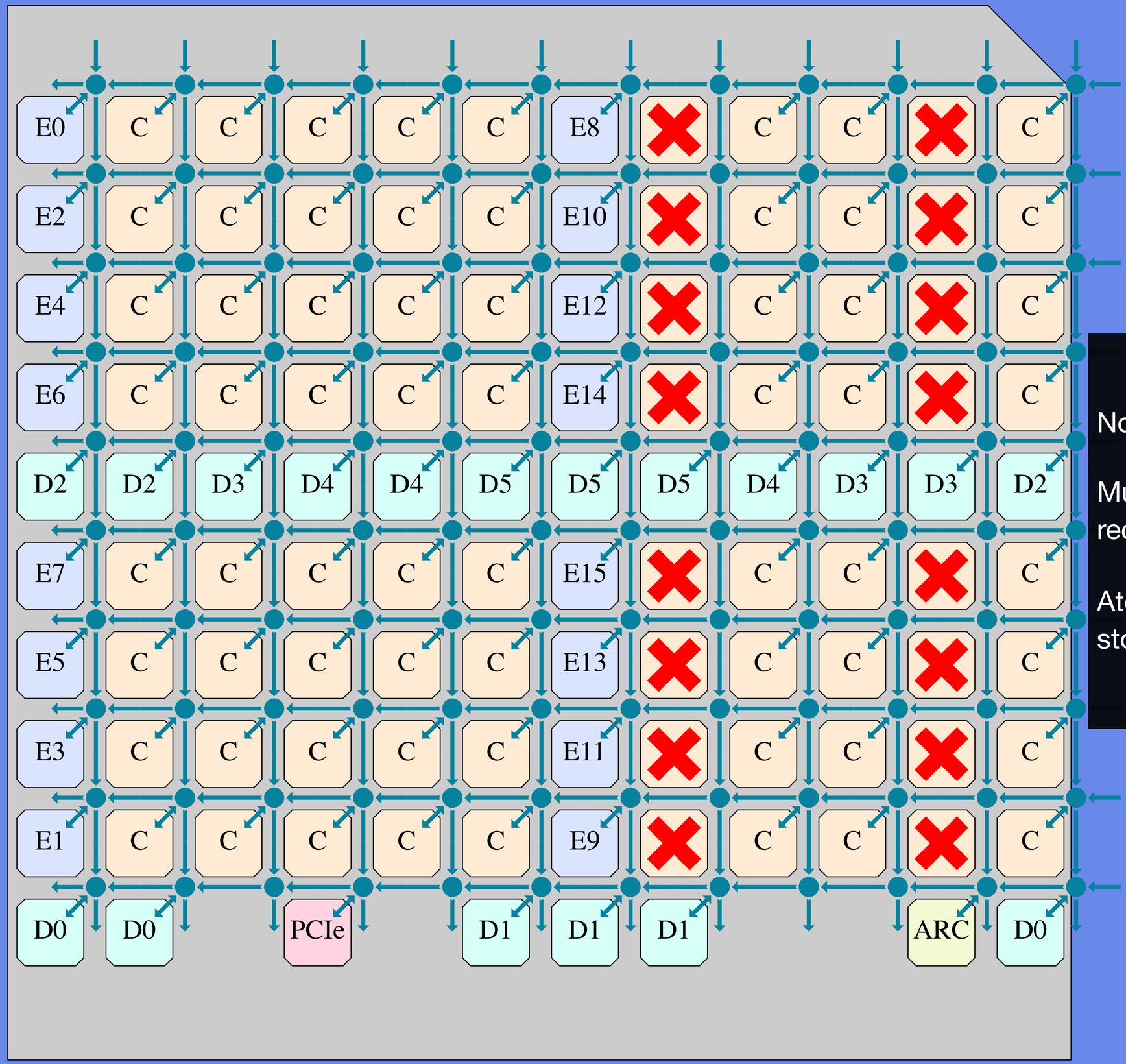
Five types of tile:

ARC (1)
C: Compute (64 or 72 or 80)
D: DRAM (18)
E: Ethernet (16)
PCIe (1)

NoC does asynchronous memcpy:

```
u4 dst_x, u4 dst_y,
u32 dst_addr,
u4 src_x, u4 src_y,
u32 src_addr,
u14 num_bytes
```

Each arrow 32 bytes wide



NoC also does:

Multicast (`dst` as an x/y/w/h rectangle rather than x/y point)

Atomics (e.g. increment at `src`, store original value to `dst`)

C C C E14 C C C
D4 D4 D5 D5 D5 D4 D3 D3 D2
C C C E15 C C C
C C C E13 C C C
C C C E11 C C C
C C C E9 C C C
PCIe D1 D1 D1 ARC D0

C E15 C C C
C E13 C C C
C E11 C C C
C E9 C C C
D1 D1 D1 ARC D0

C C C C C C C C C C C C D1 ARC D0

C

C

C

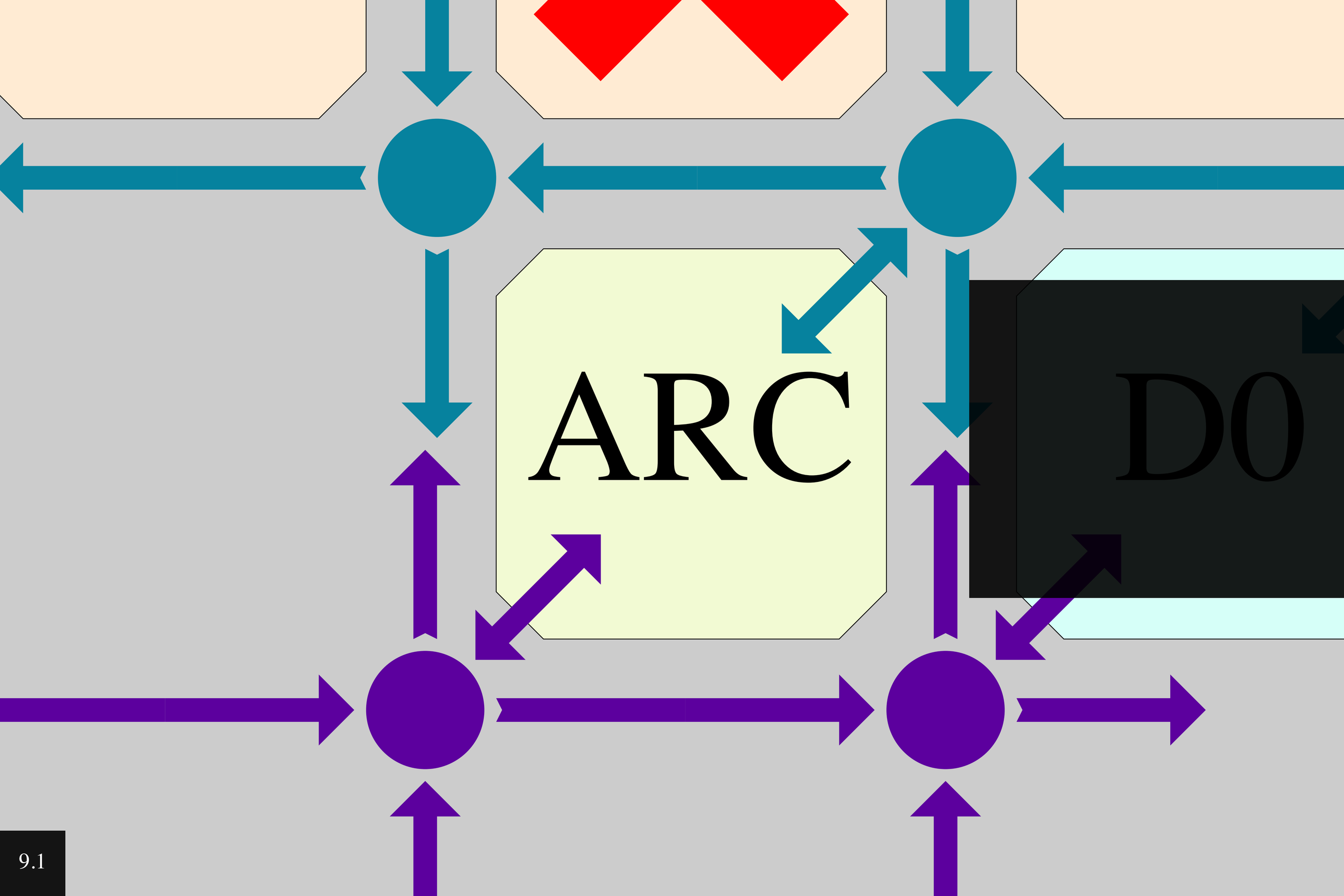
ARC

D0

ARC

ARC tile responsibilities:

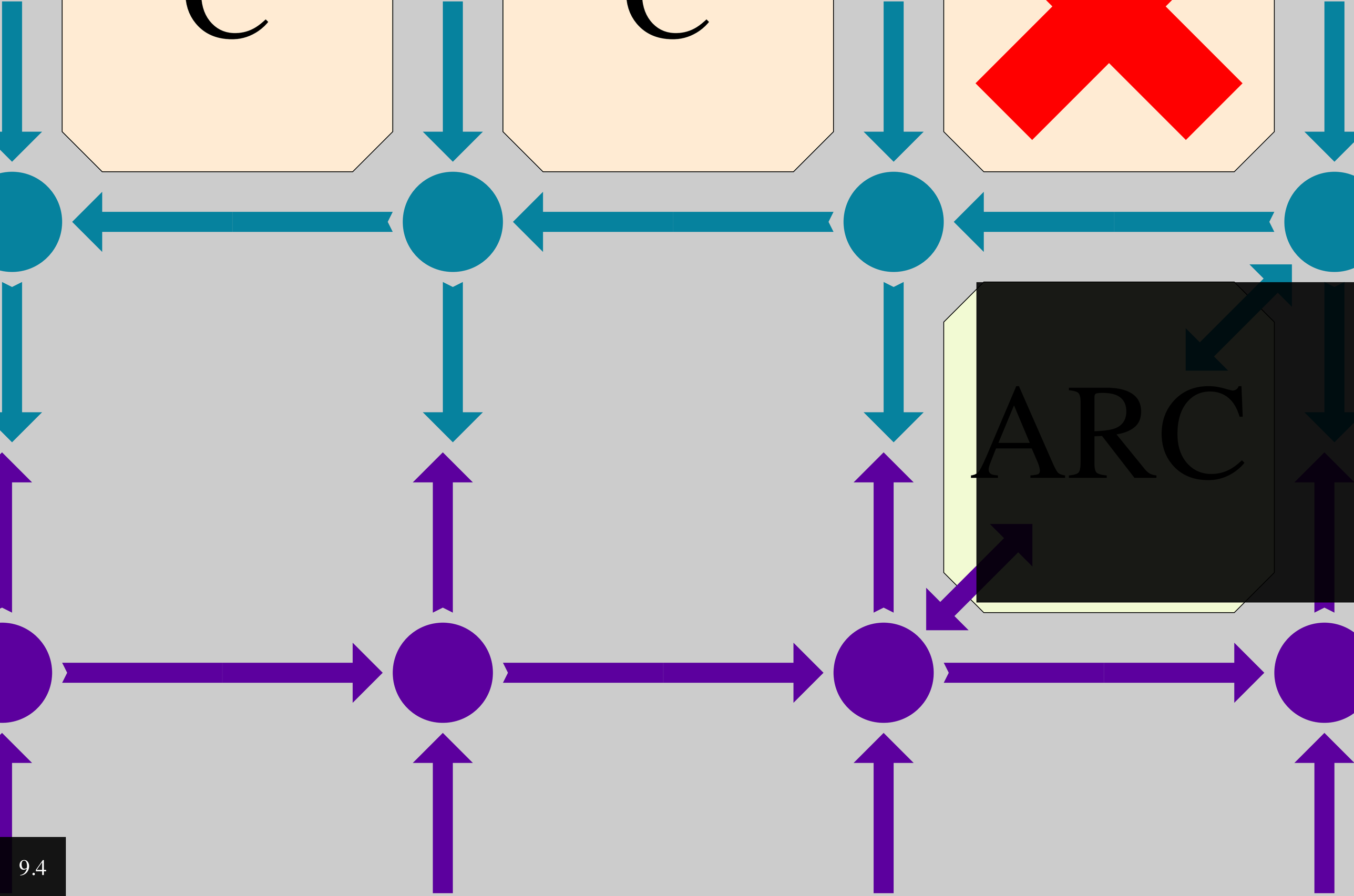
Power management
Reset management
Clock management
Thermal management

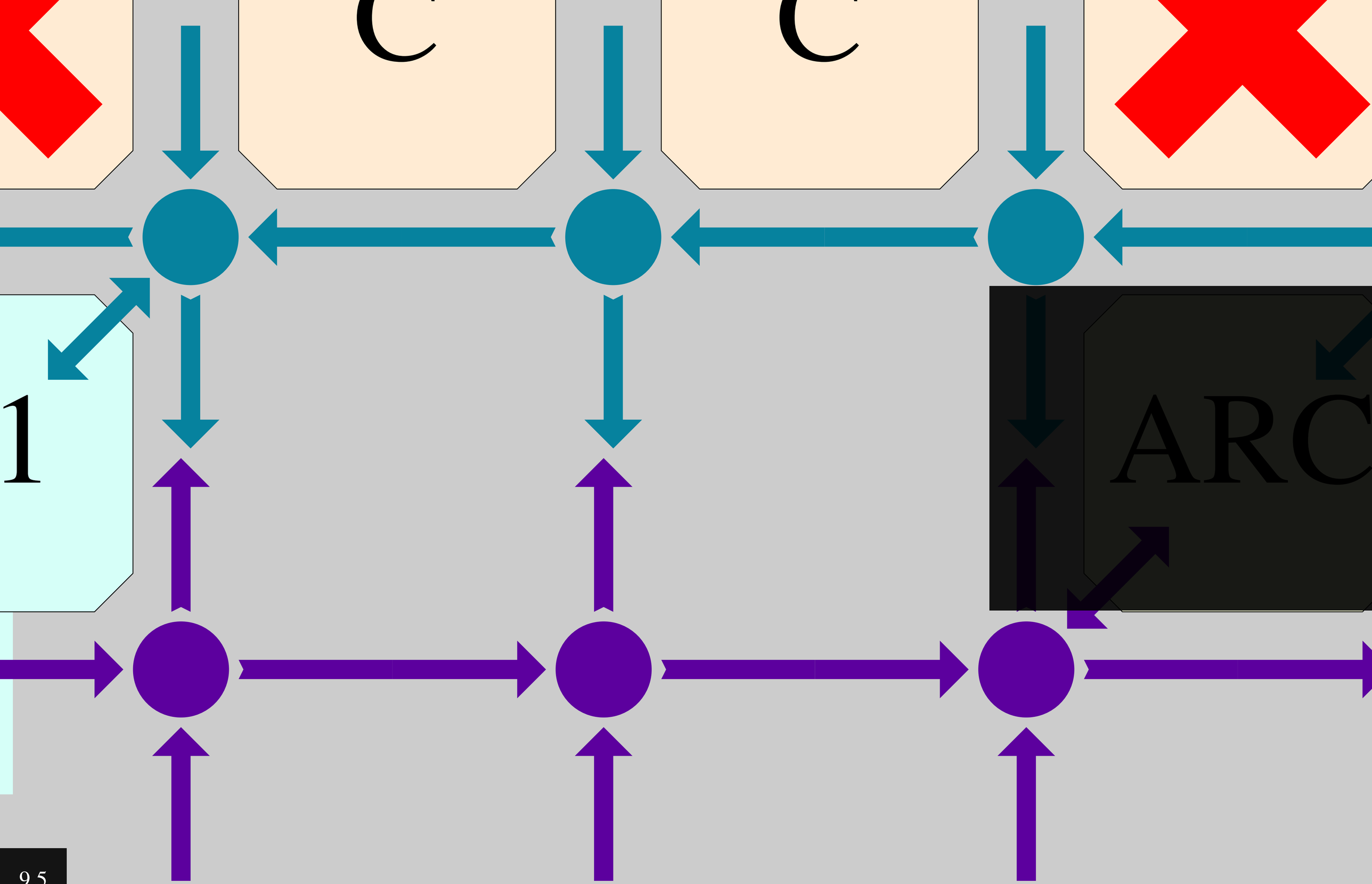

ARC
D0

ARC

D

ARC


ARC



C

C

P1

AR

C

C

P1

E9

C

C

D1

D1

E9

C

C

1

D1

D1

C

E9

C

P1

P1

P1

DRAM tile address space:

`00000000-3fffffff`: Channel 0
`40000000-7fffffff`: Channel 1
`80000000-ffffffff`: Unmapped

Each 2 GiB GDDR6 chip exposed identically on three different tiles for increased NoC interconnect

C

C

E9

D1

D1

D1

C

C

E9

e

D1

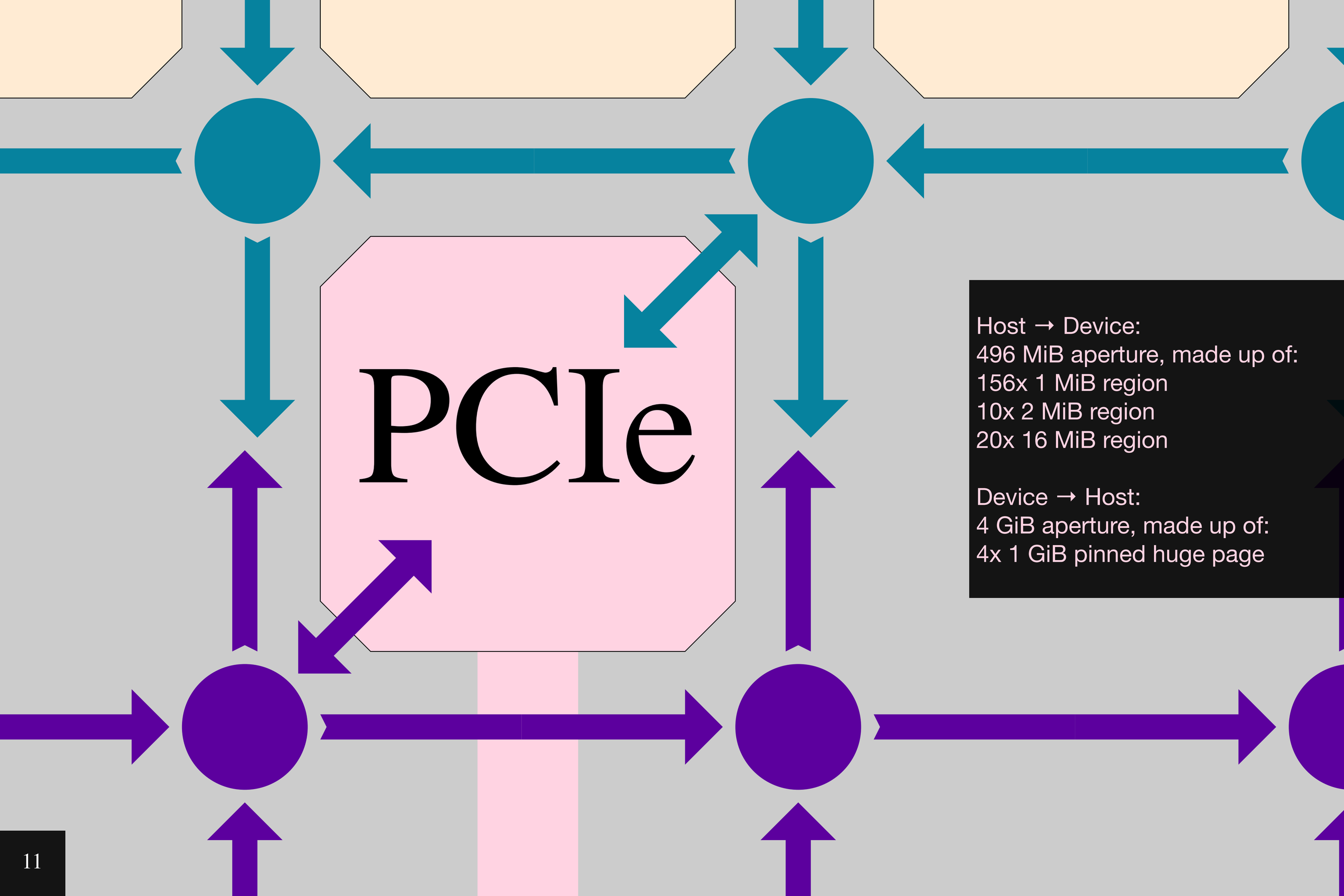
D1

PCIe

D1

PCIe

D1


PCIe
Host → Device:
496 MiB aperture, made up of:
156x 1 MiB region
10x 2 MiB region
20x 16 MiB region
Device → Host:
4 GiB aperture, made up of:
4x 1 GiB pinned huge page

PCIe

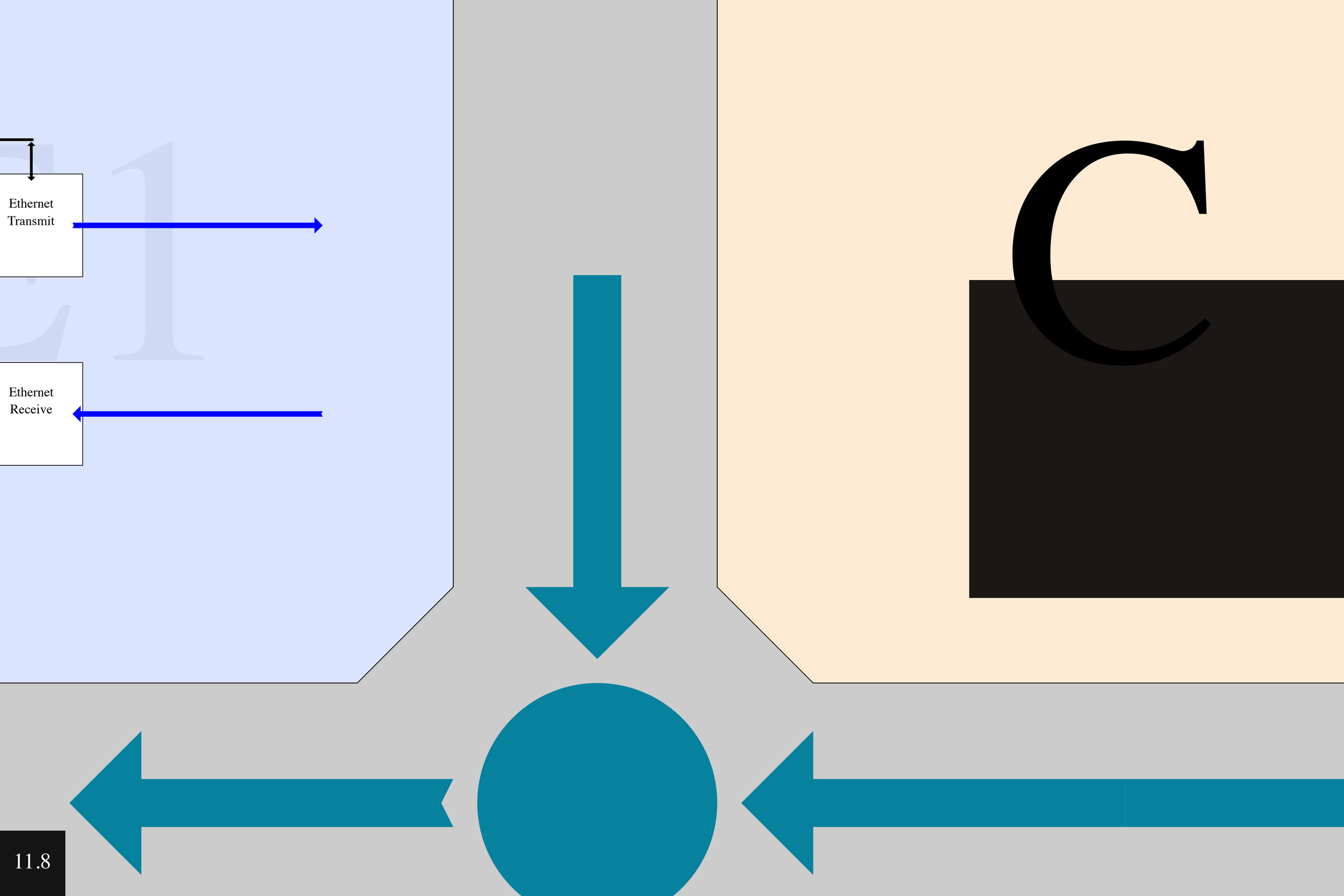
C

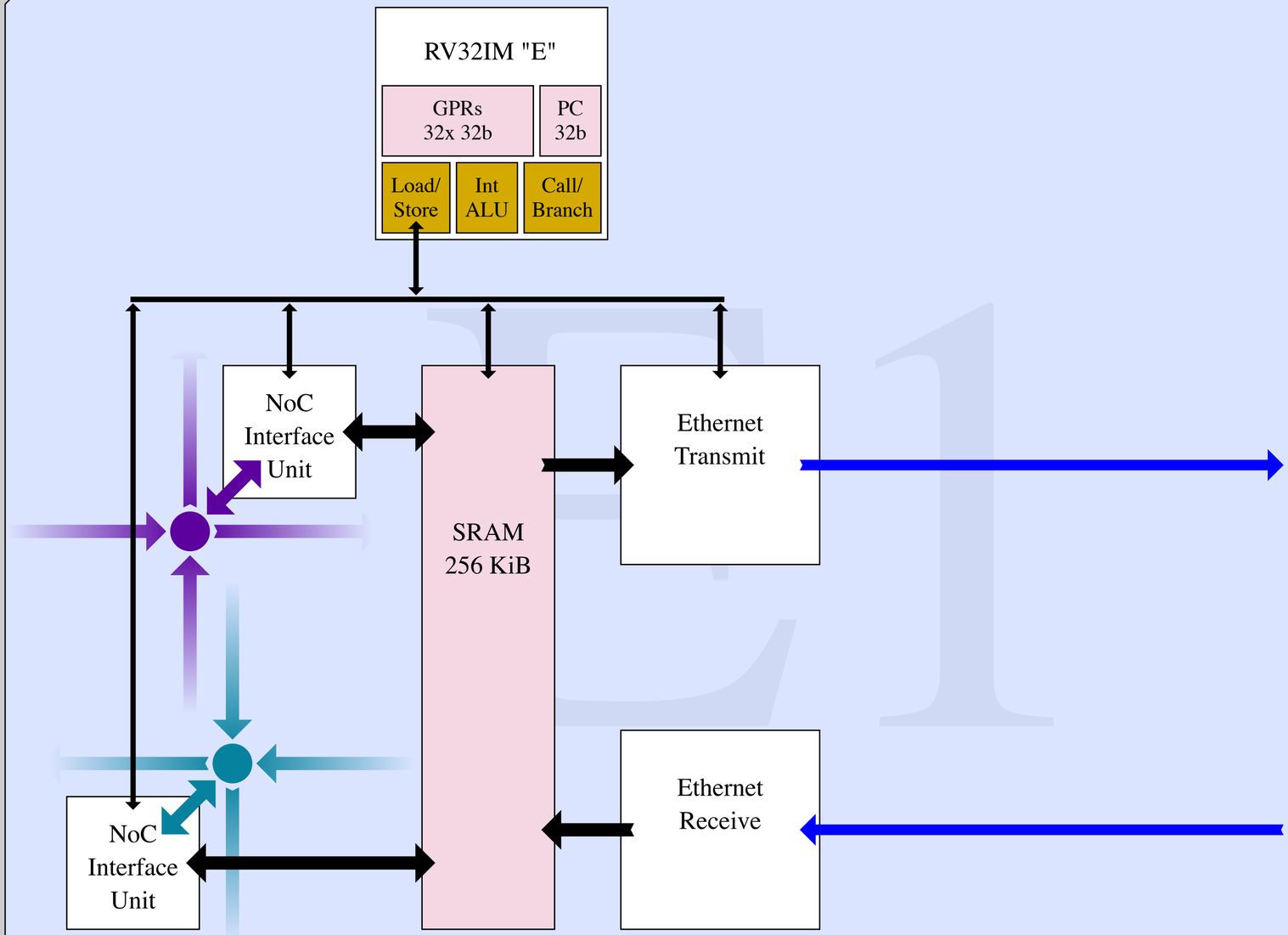
C

D0

11.4

C


Ethernet
Transmit
Ethernet
Receive



Ethernet tile address space:

`00000000-0003ffff`: SRAM
`00040000-feffffff`: Unmapped
`ff000000-ffffffff`: Peripherals

Thin arrows 4 bytes wide
Thick arrows 32 bytes wide
Blue arrows 100GbE

RV32IM "E"

GPRs
32x 32b

PC
32b

Load/
Store

Int
ALU

Call/
Branch

ce

SRAM
256 KiB

Ethernet
Transmit

Ethernet
Receive

RV32IM "NC"

GPRs
32x 32b

PC
32b

Load/
Store

Int
ALU

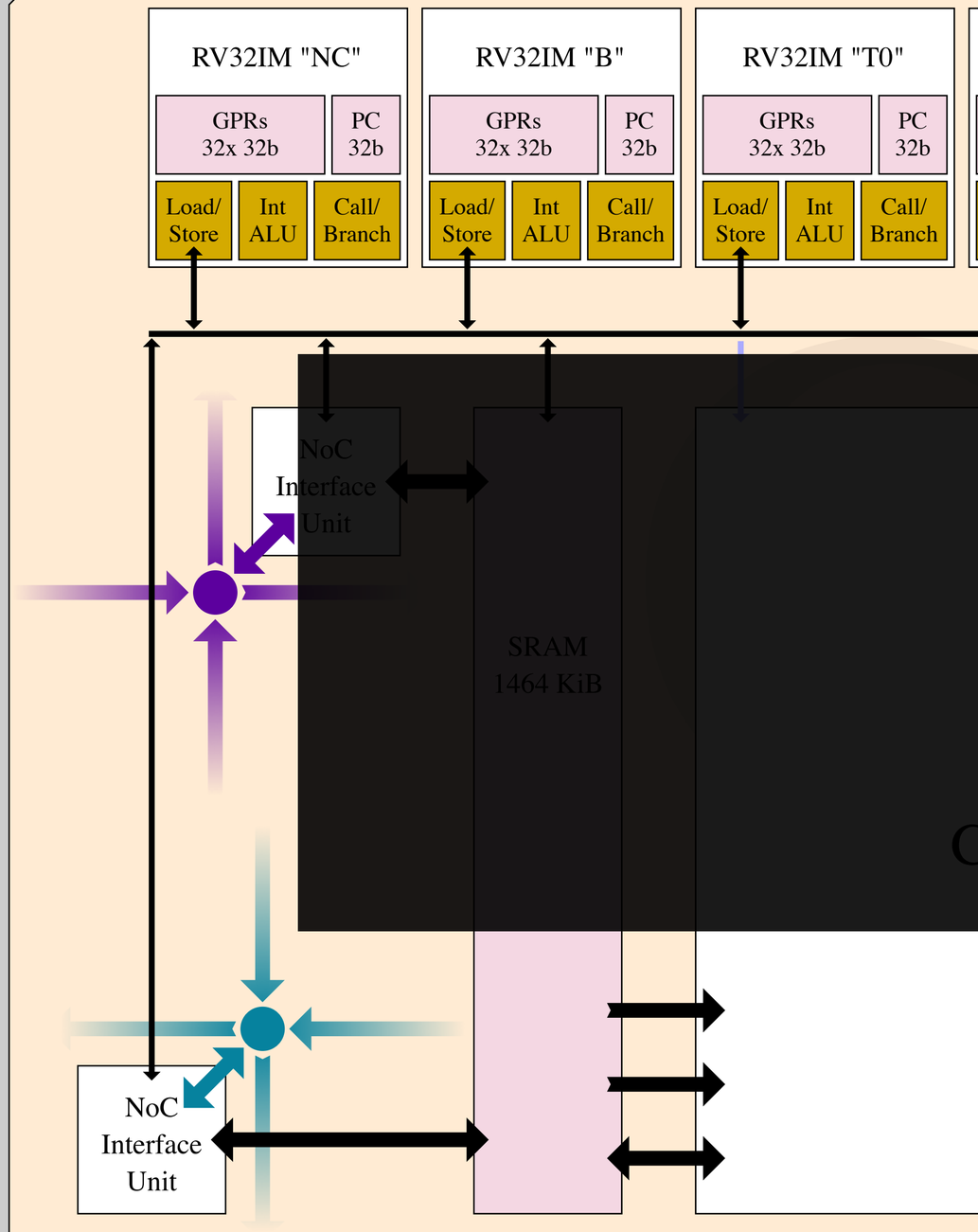
Call/
Branch

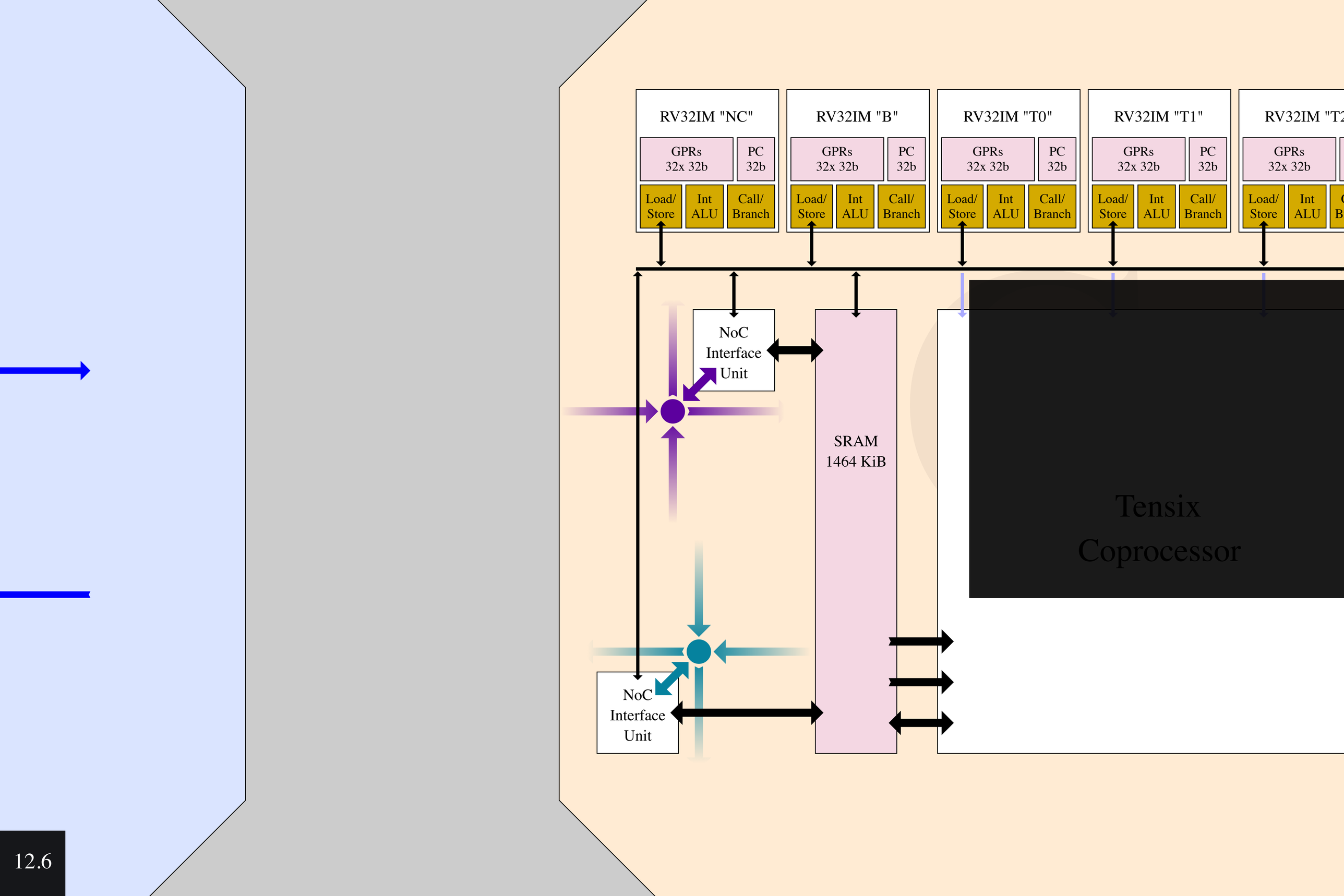
NoC
Interface
Unit

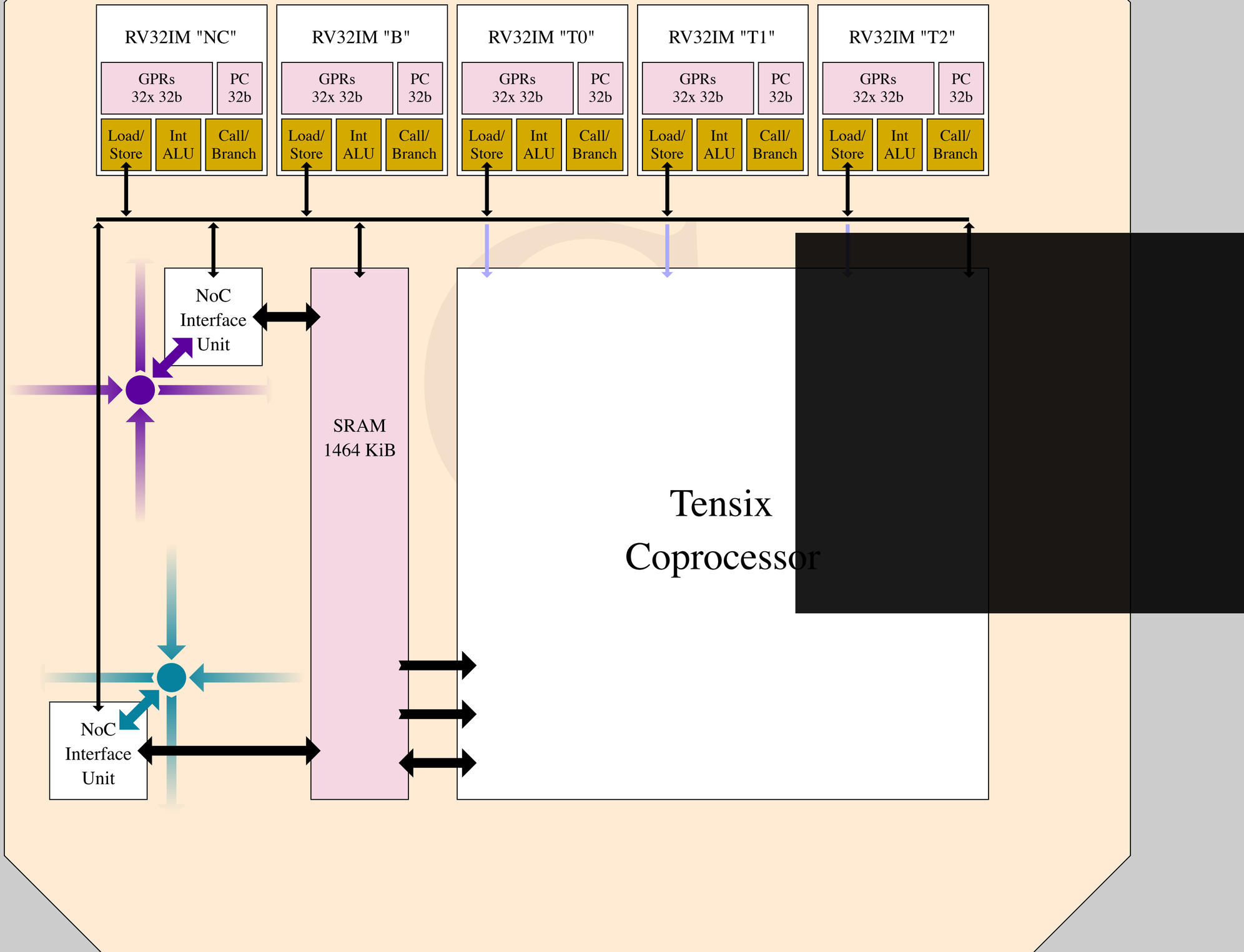
NoC
Interface
Unit

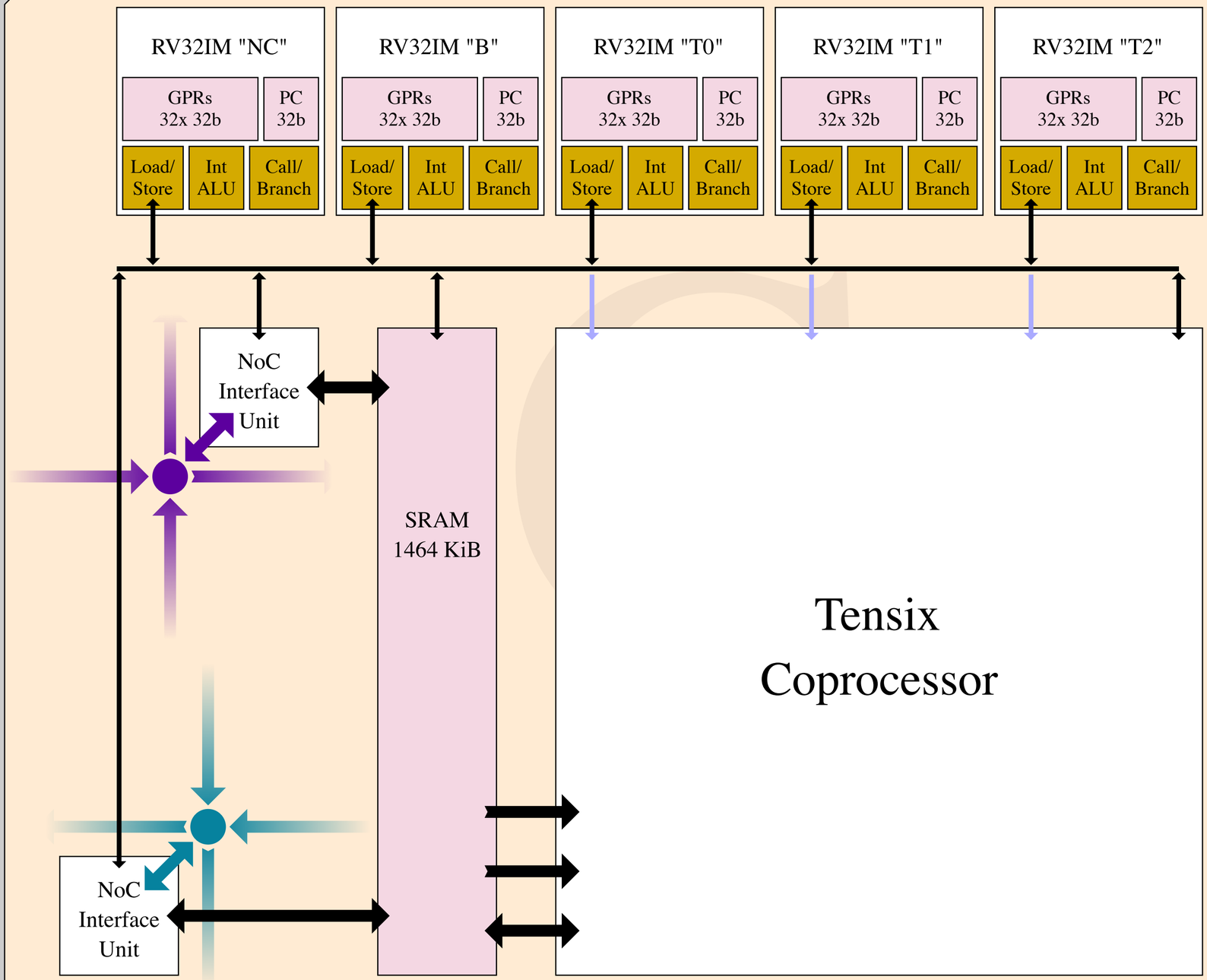
rnet
smit

rnet
eive




RV32IM "NC"
GPRs
32x 32b
PC
32b
Load/
Store
Int
ALU
Call/
Branch
RV32IM "B"
GPRs
32x 32b
PC
32b
Load/
Store
Int
ALU
Call/
Branch
RV32IM "T0"
GPRs
32x 32b
PC
32b
Load/
Store
Int
ALU
Call/
Branch
RV32IM "T1"
GPRs
32x 32b
PC
32b
Load/
Store
Int
ALU
Call/
Branch
RV32IM "T2
GPRs
32x 32b
Load/
Store
Int
ALU
NoC
Interface
Unit
SRAM
1464 KiB
Tensix
Coprocessor
NoC
Interface
Unit

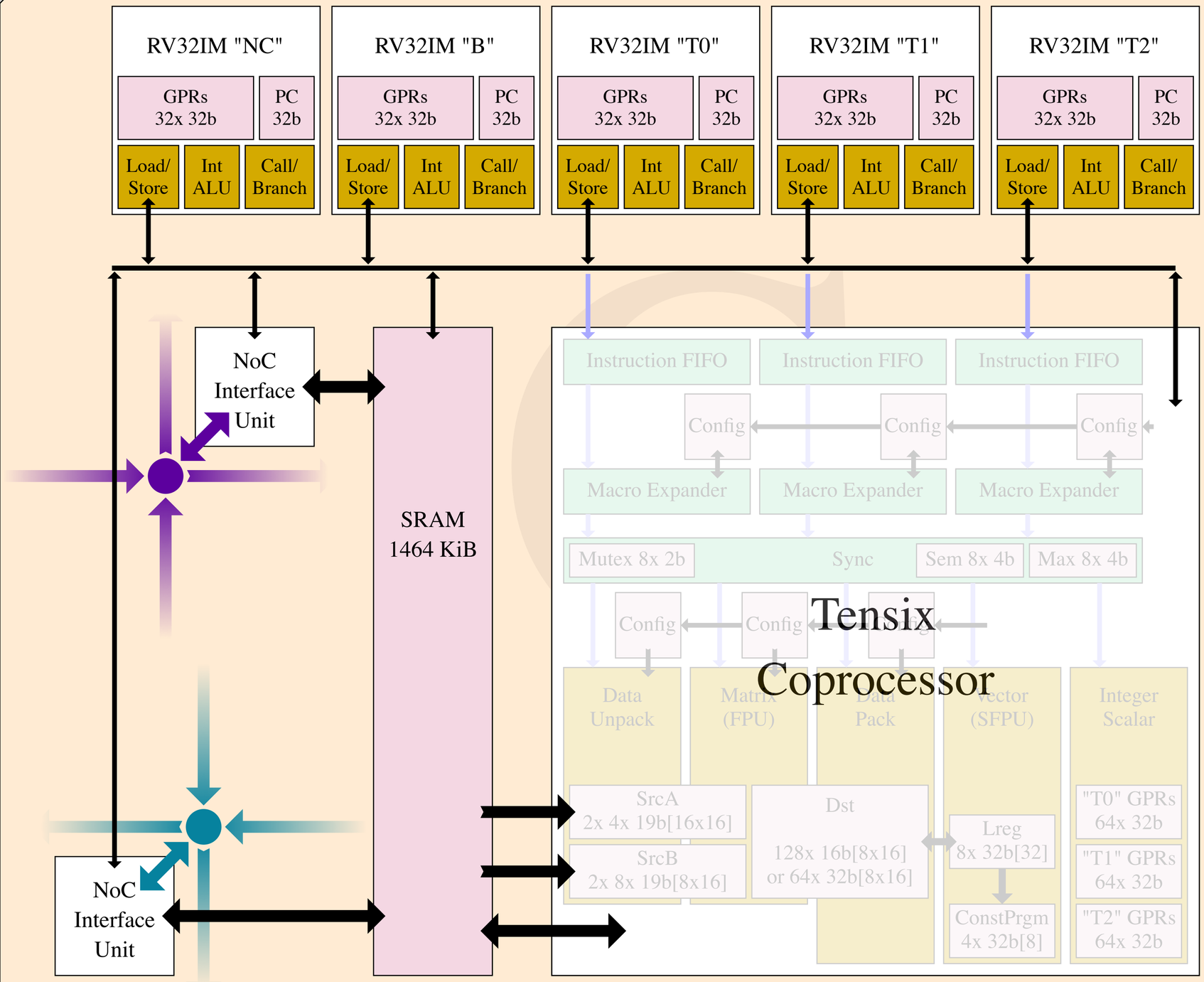

RV32IM "NC"
GPRs 32x 32b
PC 32b
Load/ Store
Int ALU
Call/ Branch
RV32IM "B"
GPRs 32x 32b
PC 32b
Load/ Store
Int ALU
Call/ Branch
RV32IM "T0"
GPRs 32x 32b
PC 32b
Load/ Store
Int ALU
Call/ Branch
RV32IM "T1"
GPRs 32x 32b
PC 32b
Load/ Store
Int ALU
Call/ Branch
RV32IM "T2"
GPRs 32x 32b
PC 32b
Load/ Store
Int ALU
Call/ Branch
NoC Interface Unit
SRAM 1464 KiB
Tensix Coprocessor
NoC Interface Unit

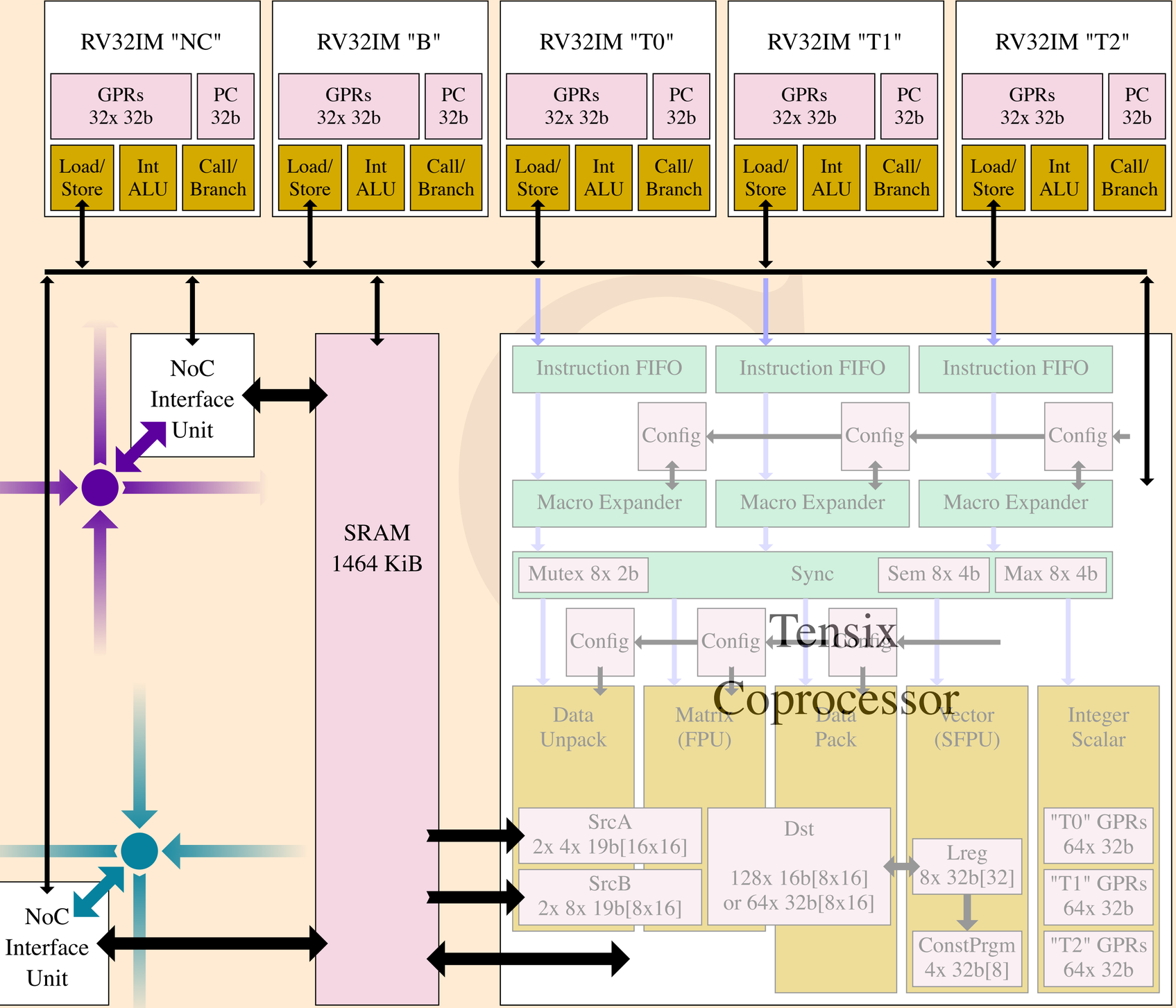


Compute tile address space:

`00000000-0016dfff`: SRAM
`0016e000-feffffff`: Unmapped
`ff000000-ffffffff`: Peripherals

Thin arrows 4 bytes wide
Thick arrows 32 bytes wide


RV32IM "NC"
RV32IM "B"
RV32IM "T0"
RV32IM "T1"
RV32IM "T2"
GPRs 32x 32b
PC 32b
Load/ Store
Int ALU
Call/ Branch
NoC Interface Unit
SRAM 1464 KiB
Instruction FIFO
Config
Macro Expander
Mutex 8x 2b
Sync
Sem 8x 4b
Max 8x 4b
Data Unpack
Matrix (FPU)
Data Pack
Vector (SFPU)
Integer Scalar
SrcA 2x 4x 19b[16x16]
SrcB 2x 8x 19b[8x16]
Dst 128x 16b[8x16] or 64x 32b[8x16]
Lreg 8x 32b[32]
ConstPrgm 4x 32b[8]
"T0" GPRs 64x 32b
"T1" GPRs 64x 32b
"T2" GPRs 64x 32b
Tensix Coprocessor


RV32IM "NC"
RV32IM "B"
RV32IM "T0"
RV32IM "T1"
RV32IM "T2"
GPRs 32x 32b
PC 32b
Load/ Store
Int ALU
Call/ Branch
NoC Interface Unit
SRAM 1464 KiB
Instruction FIFO
Config
Macro Expander
Mutex 8x 2b
Sync
Sem 8x 4b
Max 8x 4b
Tensix Coprocessor
Data Unpack
Matrix (FPU)
Data Pack
Vector (SFPU)
Integer Scalar
SrcA 2x 4x 19b[16x16]
SrcB 2x 8x 19b[8x16]
Dst 128x 16b[8x16] or 64x 32b[8x16]
Lreg 8x 32b[32]
ConstPrgm 4x 32b[8]
"T0" GPRs 64x 32b
"T1" GPRs 64x 32b
"T2" GPRs 64x 32b

GPRs 32x 32b | PC 32b | Load/ Store | Int ALU | Call/ Branch

NoC Interface Unit

SRAM 1464 KiB

Instruction FIFO | Config | Macro Expander

Mutex 8x 2b | Sync | Sem 8x 4b | Max 8x 4b

Tensix Coprocessor

Data Unpack | Matrix (FPU) | Data Pack | Vector (SFPU) | Integer Scalar

SrcA 2x 4x 19b[16x16]

SrcB 2x 8x 19b[8x16]

Dst 128x 16b[8x16] or 64x 32b[8x16]

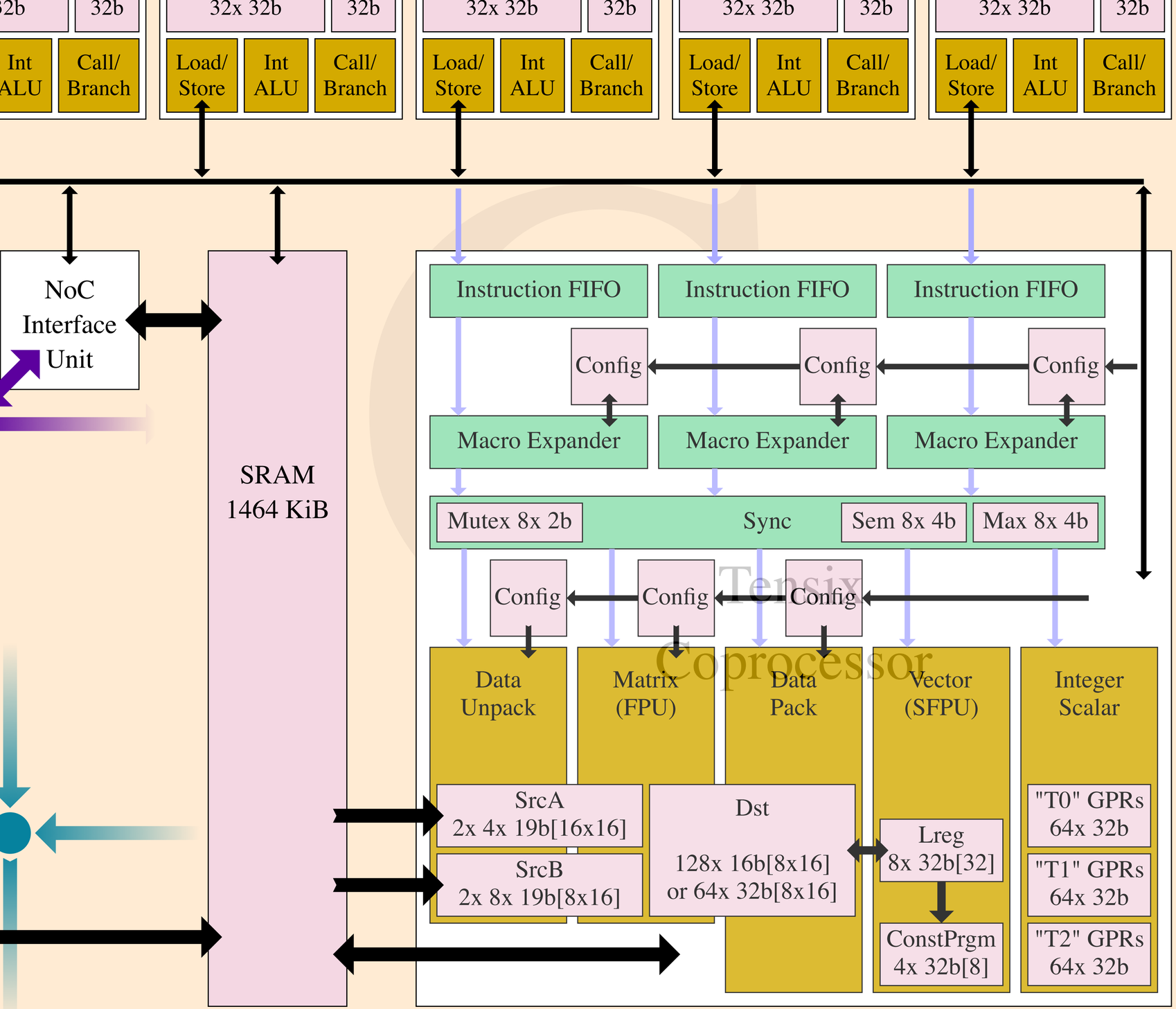
Lreg 8x 32b[32]

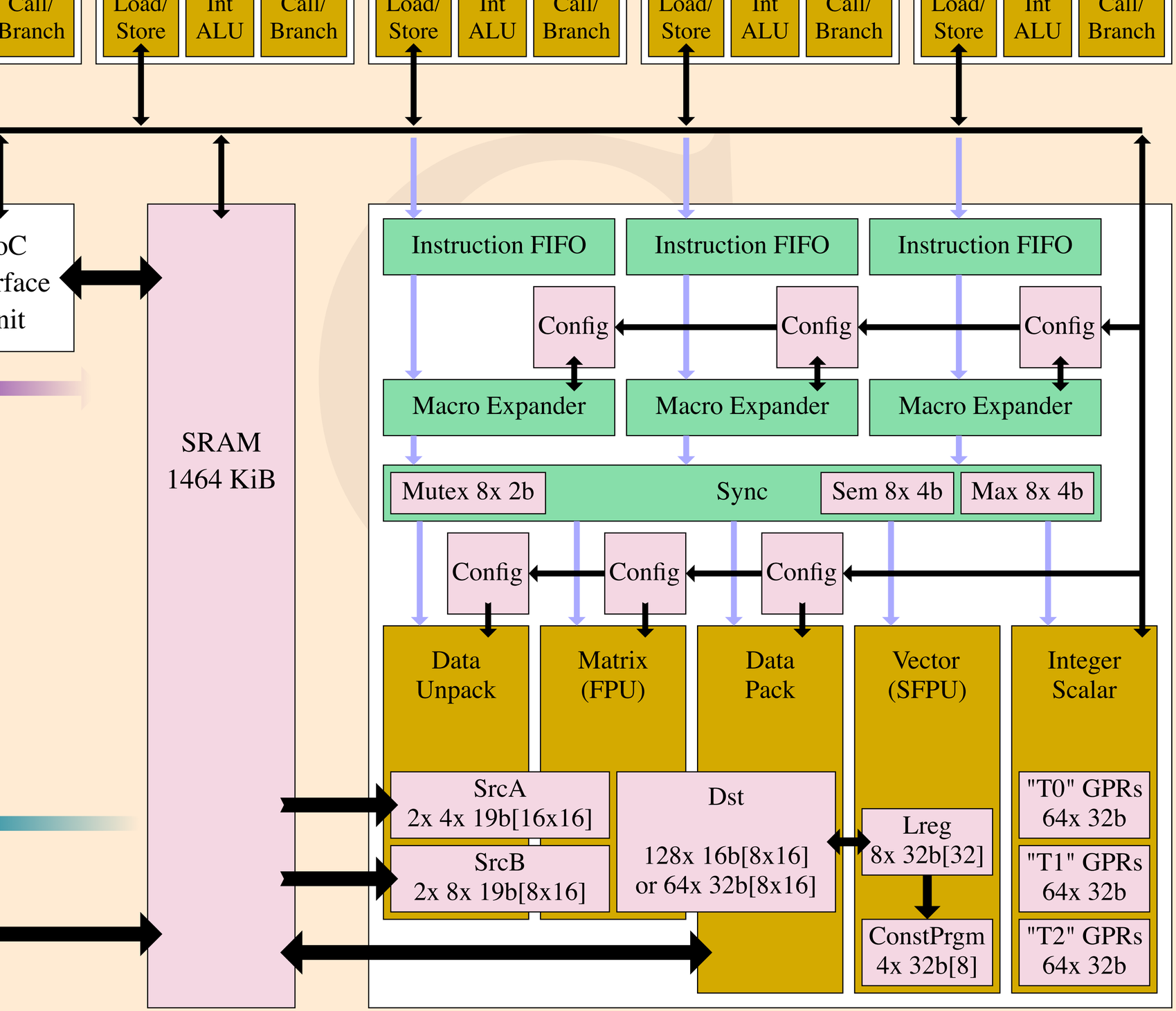
ConstPrgm 4x 32b[8]

"T0" GPRs 64x 32b

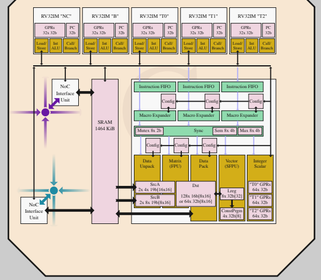
"T1" GPRs 64x 32b

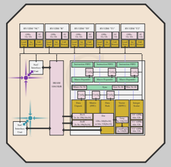
"T2" GPRs 64x 32b


32b
32b
Int ALU
Call/ Branch
32x 32b
32b
Load/ Store
Int ALU
Call/ Branch
32x 32b
32b
Load/ Store
Int ALU
Call/ Branch
32x 32b
32b
Load/ Store
Int ALU
Call/ Branch
32x 32b
32b
Load/ Store
Int ALU
Call/ Branch
NoC Interface Unit
SRAM 1464 KiB
Instruction FIFO
Instruction FIFO
Instruction FIFO
Config
Config
Config
Macro Expander
Macro Expander
Macro Expander
Mutex 8x 2b
Sync
Sem 8x 4b
Max 8x 4b
Config
Config
Config
Tensix
Coprocessor
Data Unpack
Matrix (FPU)
Data Pack
Vector (SFPU)
Integer Scalar
SrcA 2x 4x 19b[16x16]
SrcB 2x 8x 19b[8x16]
Dst 128x 16b[8x16] or 64x 32b[8x16]
Lreg 8x 32b[32]
ConstPrgm 4x 32b[8]
"T0" GPRs 64x 32b
"T1" GPRs 64x 32b
"T2" GPRs 64x 32b



Data Unpack: SRAM → SrcA
Data Unpack: SRAM → SrcB
Matrix: Dst += SrcB @ SrcA
Matrix: 8x16 += 8x16 @ 16x16
Matrix: fp8, ½ rate bf16, ¼ rate fp16
Data Pack: Dst → SRAM
Data Pack: or SRAM += Dst
Vector: 32 lanes of SIMD
Vector: fp32 or int32

E2	C	C	✕	C	C
E1	C	C	✕	C	C
E0		C	✕	C	C
D0	D0	D0	ARC		

E4	C	C	✕	C	C	C	C	C
E3	C	C	✕	C	C	C	C	C
E2	C	C	✕	C	C	C	C	C
E1	C	C	✕	C	C	C	C	C
E0		C	✕	C	C	C	C	C
D0	D0	D0	ARC			PCIe		

14.3

2 GiB

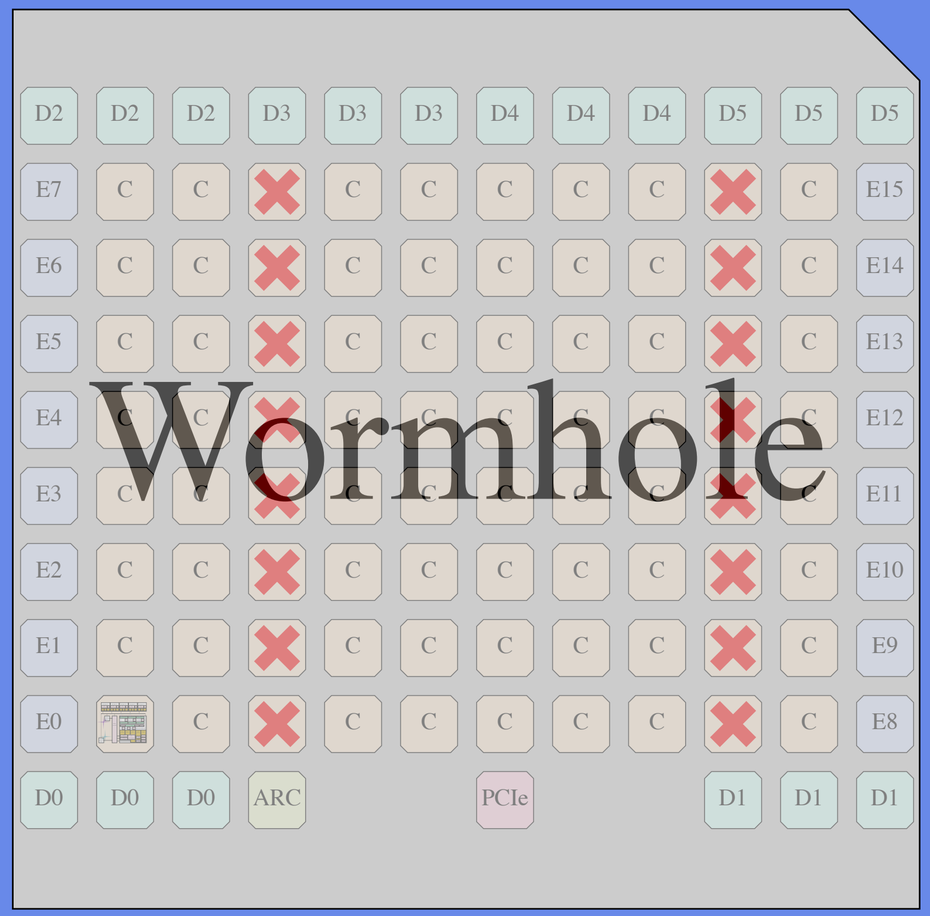
2 GiB

2 GiB GDDR6

2 GiB GDDR6

GDDR6 GDDR6 GDDR6 GDDR6 GDDR6

Cage



Cage

2 GiB GDDR6

2 GiB GDDR6

2 G GD

14.5

2 GiB GDDR6

2 GiB GDDR6

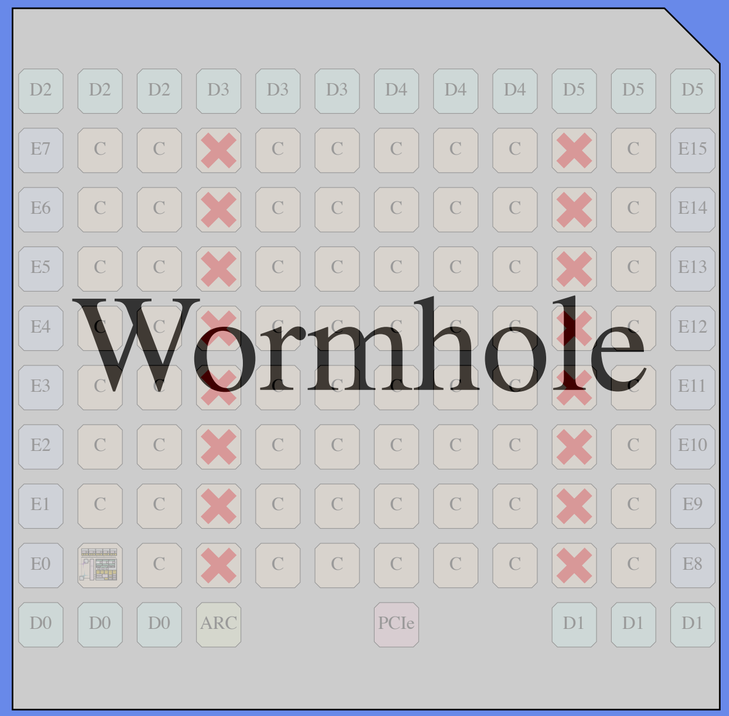
2 GiB GDDR6

2 GiB GDDR6

2 GiB GDDR6

2 GiB GDDR6

P-DD Cage



Wormhole

Worm

P-DD Cage

2 GiB GDDR6

2 GiB GDDR6

2 GiB GDDR6

14.6

Warp 100 Warp 100

2 GiB GDDR6 | 2 GiB GDDR6 | 2 GiB GDDR6 | 2 GiB GDDR6 | 2 GiB GDDR6 | 2 GiB GDDR6 | 2 GiB GDDR6 | 2 Gi GDD

QSFP-DD Cage

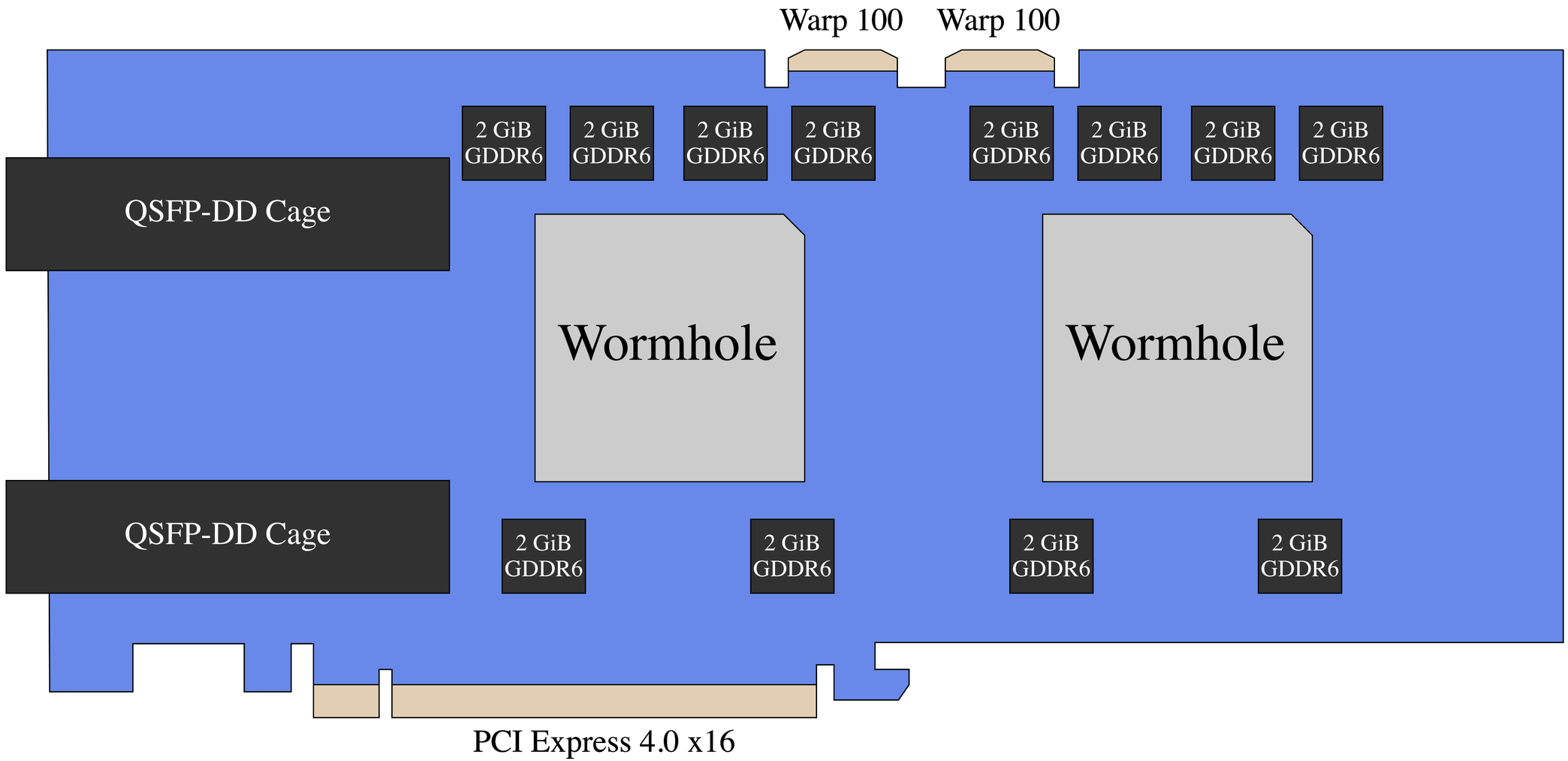
Wormhole

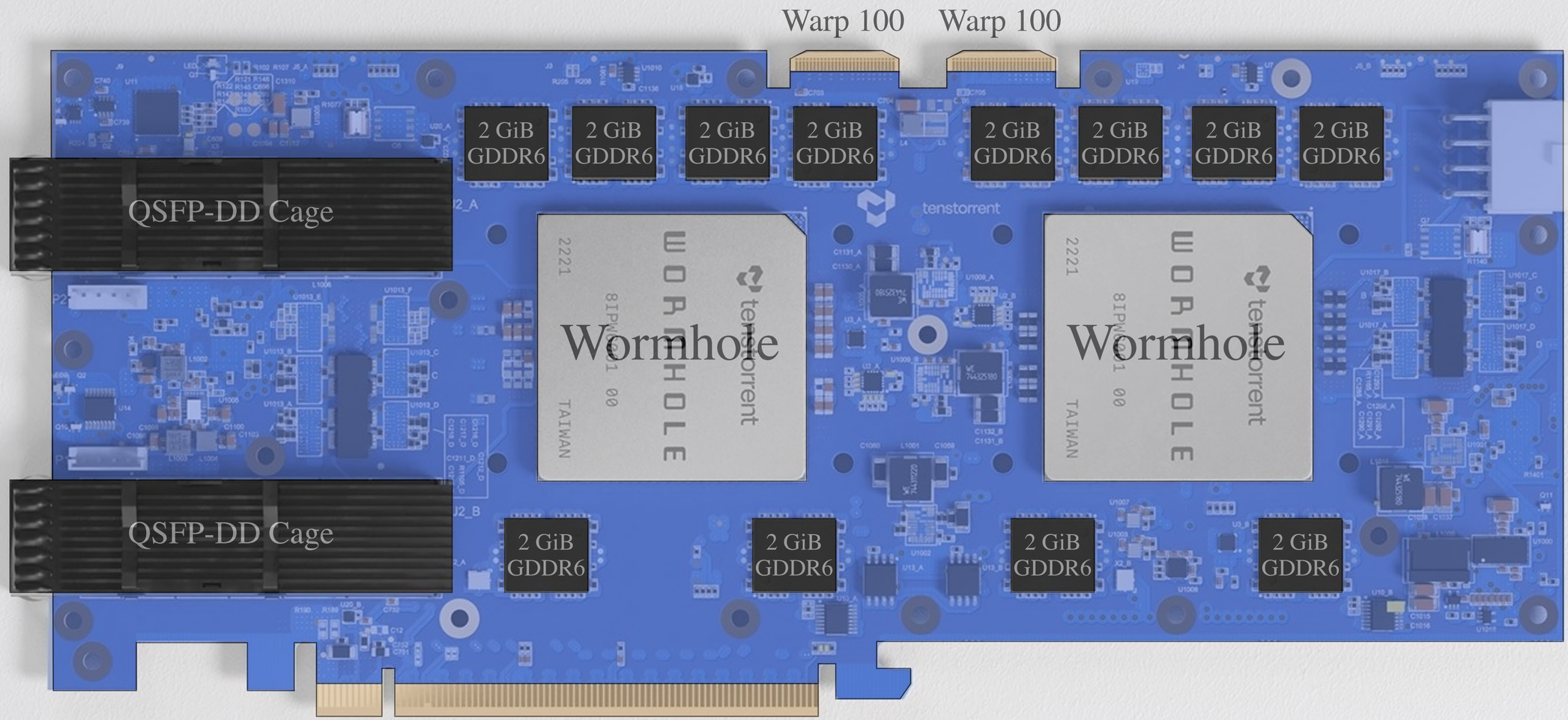
Wormhole

QSFP-DD Cage

2 GiB GDDR6 | 2 GiB GDDR6 | 2 GiB GDDR6 | 2 GiB GDDR6

PCI Express 4.0 x16


Warp 100
Warp 100
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
QSFP-DD Cage
Wormhole
Wormhole
QSFP-DD Cage
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
PCI Express 4.0 x16


Warp 100
Warp 100
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
QSFP-DD Cage
Wormhole
Wormhole
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
2 GiB GDDR6
QSFP-DD Cage
PCI Express 4.0 x16


tenstorrent
WORMHOLE
8IPW6001 00
2221 TAIWAN
tenstorrent
WORMHOLE
8IPW6001 00
2221 TAIWAN