Supermeshes

A New Network Architecture for AI Datacenters

Bill McColl

AI Compute

Deep learning, latent diffusion, transformers, flash attention, KV compression, numerically stable quantization, sparsification, pipeline scheduling, model distillation, Mixture-of-Experts, reinforcement learning, group relative policy optimization, chain-of-thought, reasoning, Monte Carlo tree search, beam search, shared knowledge, RAG, multi-modal, agents,…

Scaling laws for pre-training, post-training, and test-time compute

Fast (System 1) responses and slower (System 2) reasoning, with variable compute for complex reasoning

AI Architectures 2010-2024

- Arithmetic dominates communication (high arithmetic intensity)
- Optimized for simple dense regular computations (GEMMs): GPUs, Tensor Cores
- Low-degree uniform HPC interconnects - assume cost and speed of links is independent of distance!

AI Architecture Requirements 2025-

- Communication dominates arithmetic (low arithmetic intensity)
- Optimized for complex sparse irregular computations (also in HPC, graph computing, data science)
- Support for huge numbers of memory-bound and communication-bound tasks running concurrently (some running on a single node for less than a second, others running on thousands of nodes for hours)
- Hierarchically scalable (SoCs, chiplets, wafers, personal AI devices, servers, racks, supernodes, large clusters, massive datacenters and factories)
- Hierarchical interconnect (links may be many orders of magnitude different in cost, speed and distance)

Supermeshes

Hierarchical Localized Interconnect

- ✓ Dense at short-reach, sparse at long-reach. Best choices at each scale level for cost, speed and distance requirements
- ✓ Virtualizable and decomposable, with full network isolation ensuring zero congestion between apps for predictability and flexible optimized sharing of resources

Balanced and Data-Centric

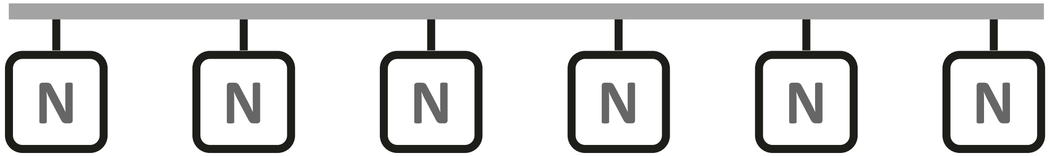
- ✓ Optimized for apps that are memory-bound and communication-bound, requiring rounds of massive data exchanges
- ✓ Scalable algorithms, communications and software optimized for hierarchical connectivity – use few hops, use all links

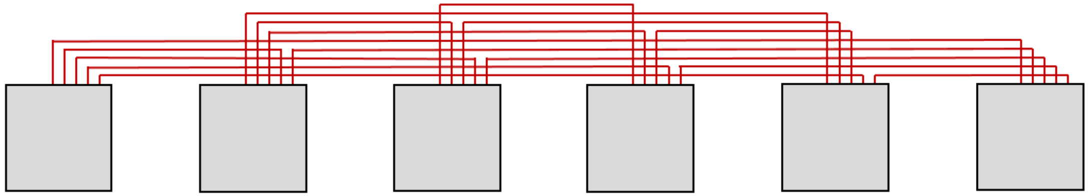
Supermeshes - Better, Faster <u>and Cheaper</u> Architectures

System Cost	
Memory	$$$$$
Compute	$$$$
Network	$$

Supermesh interconnects may increase network cost slightly, but can reduce latency and increase bandwidth significantly, enabling optimal performance and utilization of the much more costly compute nodes and memory systems

1D Supermesh SM(6)





2D Layout

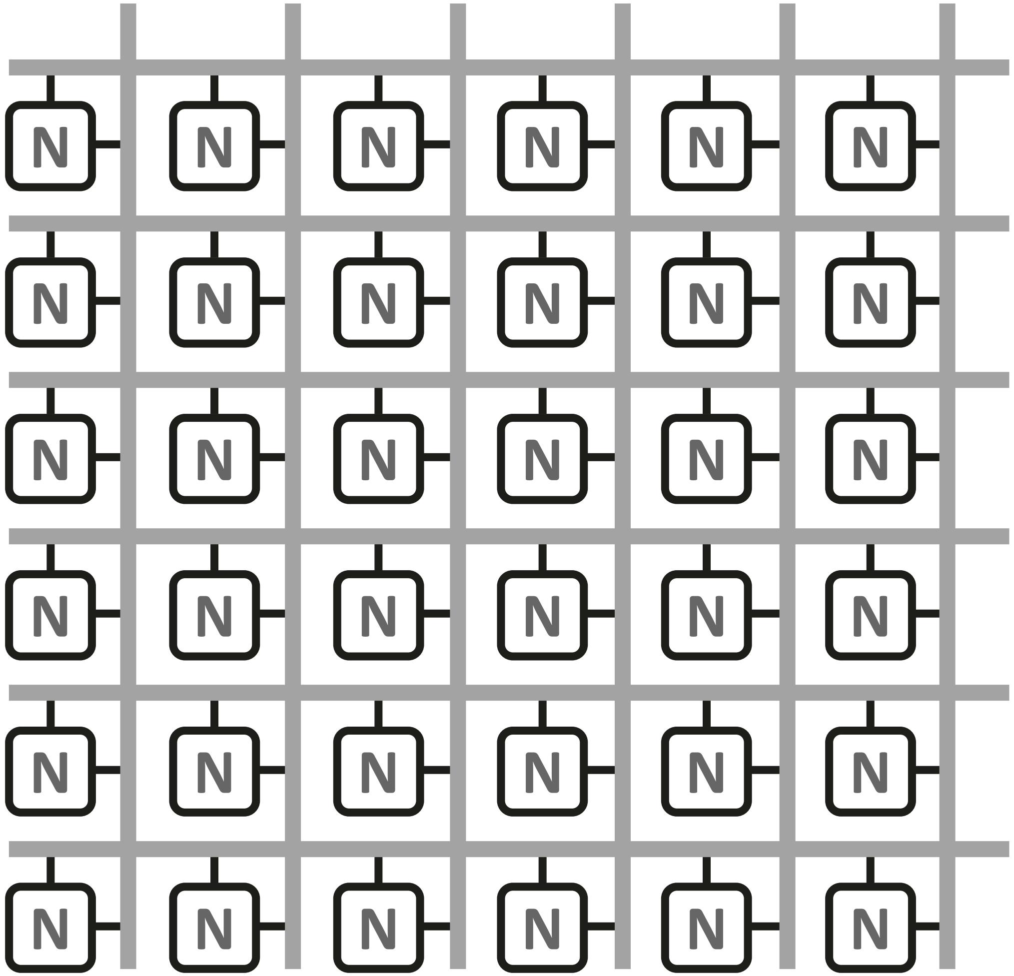
Fully connected. Single hop communication

Number of short-reach links = $\binom{m}{2}$

Bisection width $\approx m^2/4$

Minimal latency
Maximum bandwidth

2D Supermesh SM(6,6)



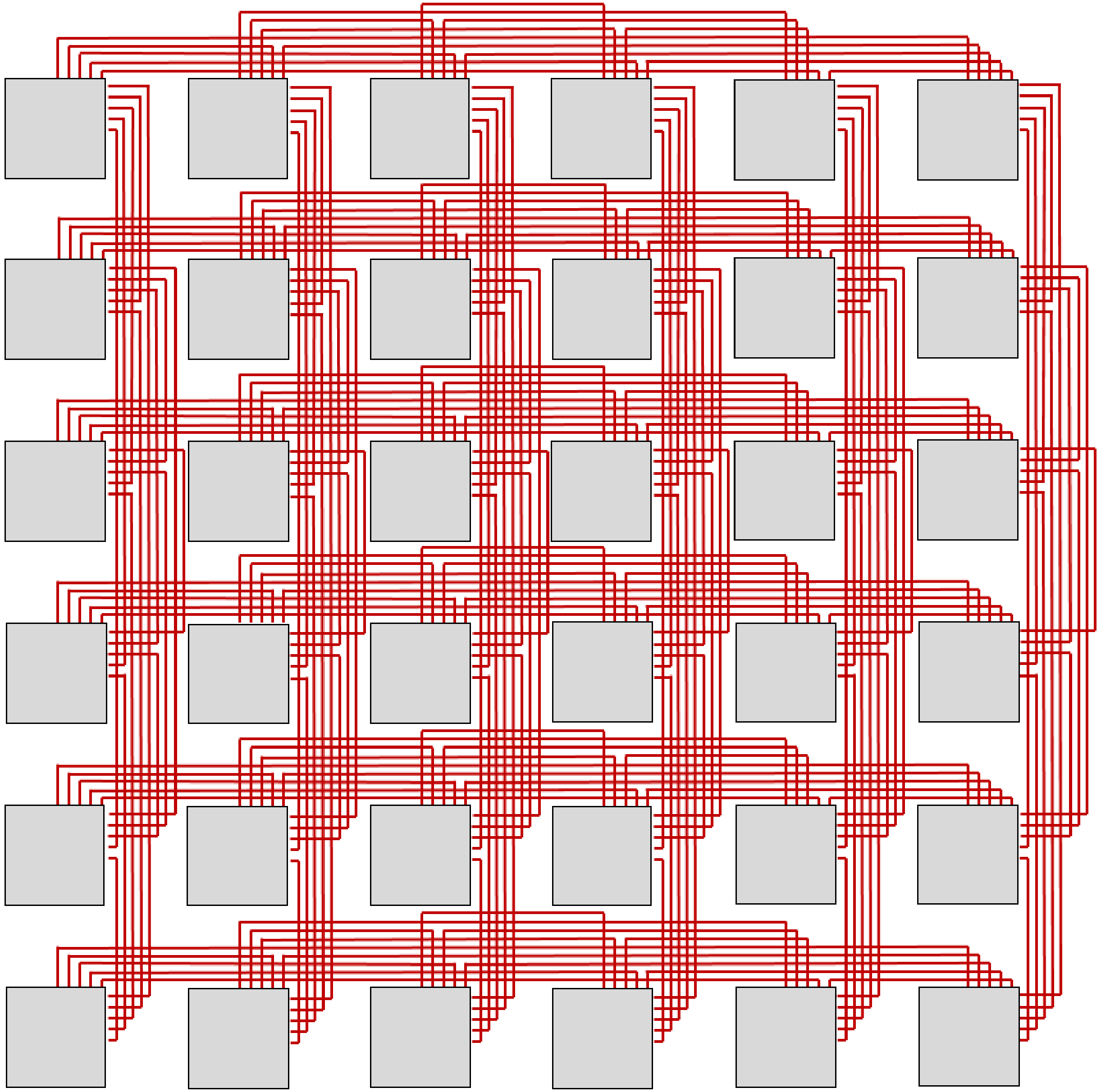
Max 2-hop communication

Number of short-reach links = $n\binom{m}{2} + m\binom{n}{2}$

Bisection width ≈ min { $nm^2/4$ **,** $mn^2/4$ **}**

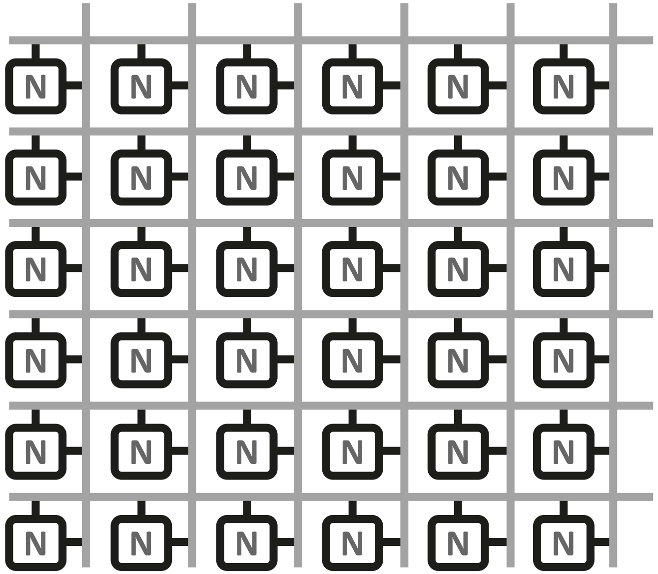
Ultra-low latency
Ultra-high bandwidth

2D Supermesh SM(6,6)

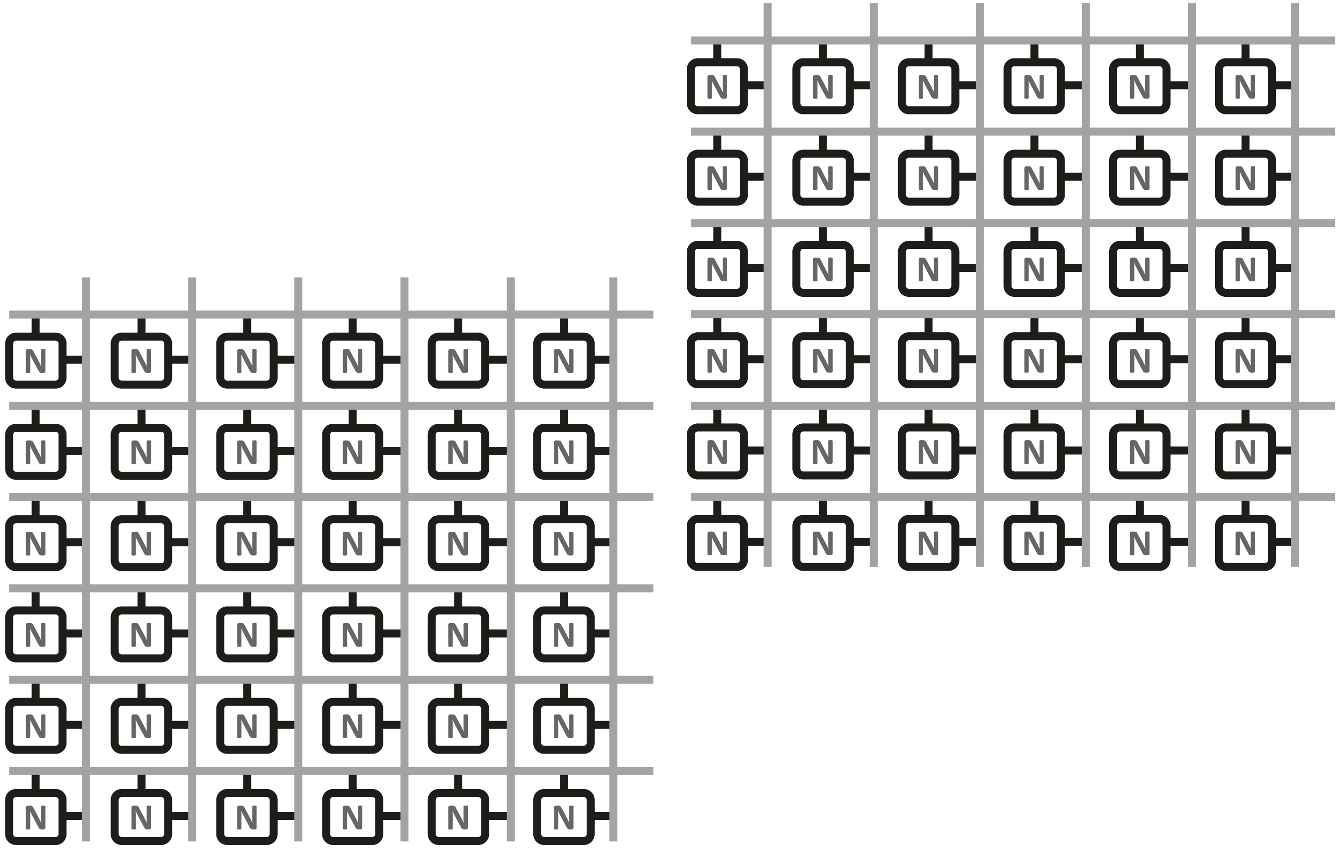


2D Layout

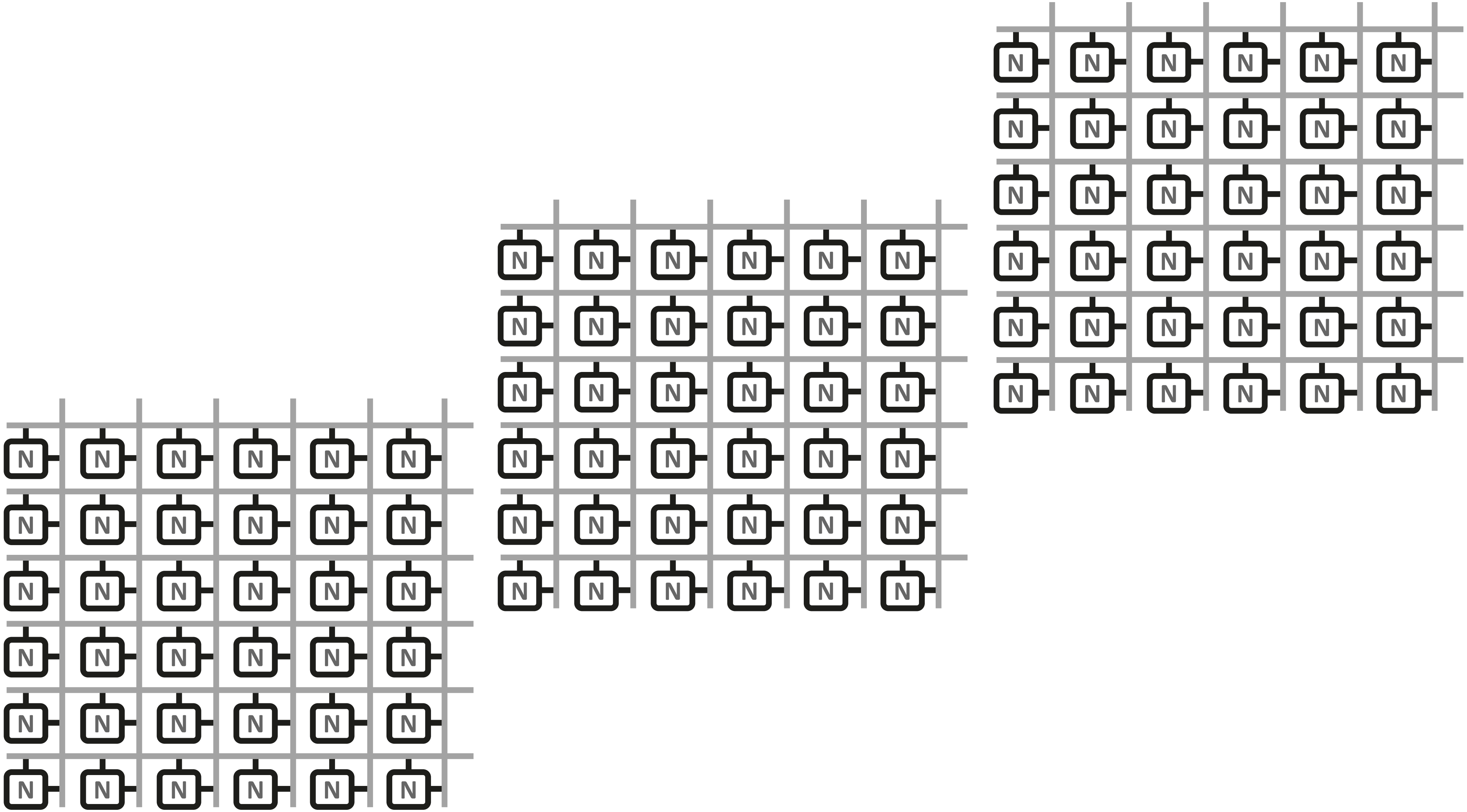
2D -> 3D



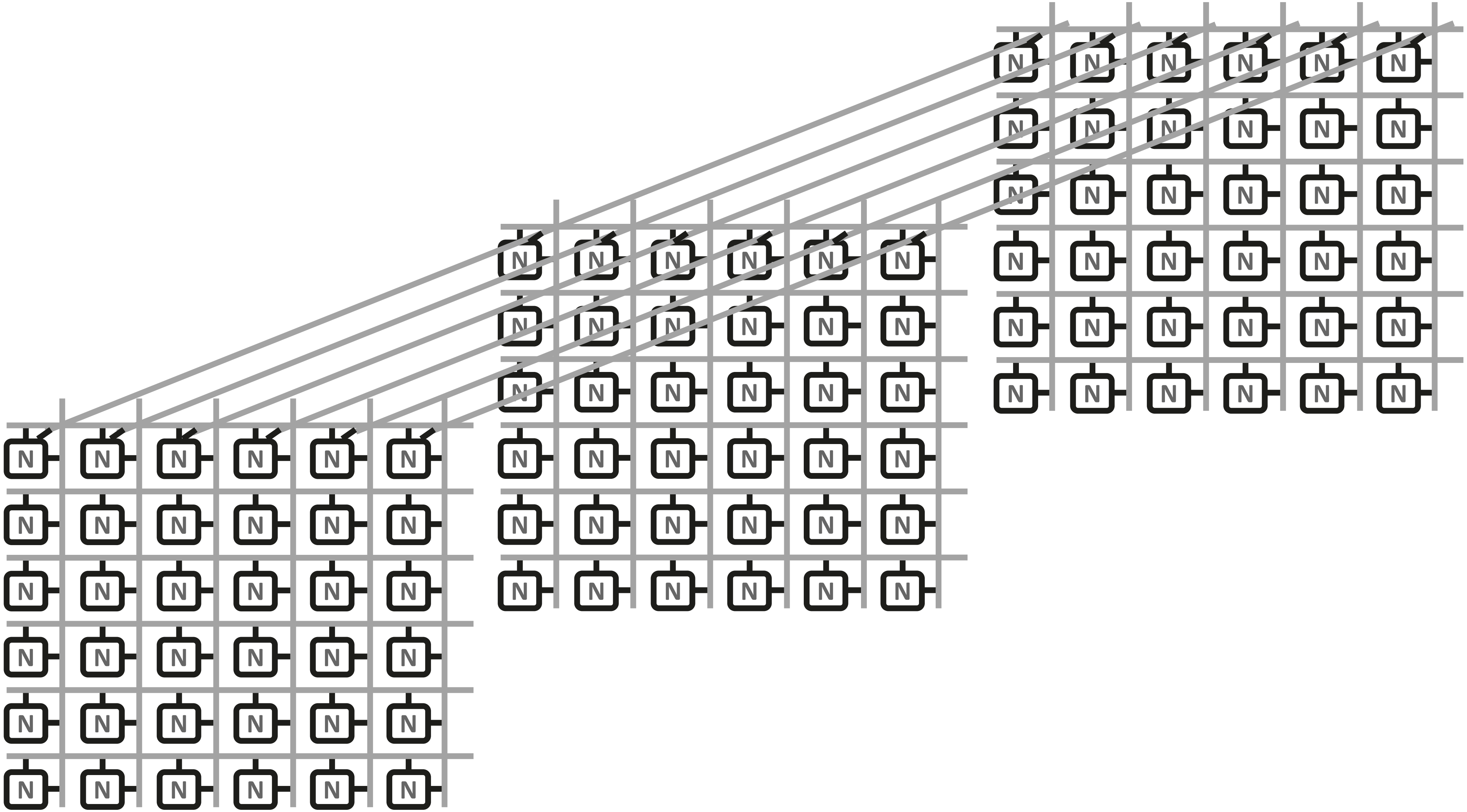
2D -> 3D



2D -> 3D

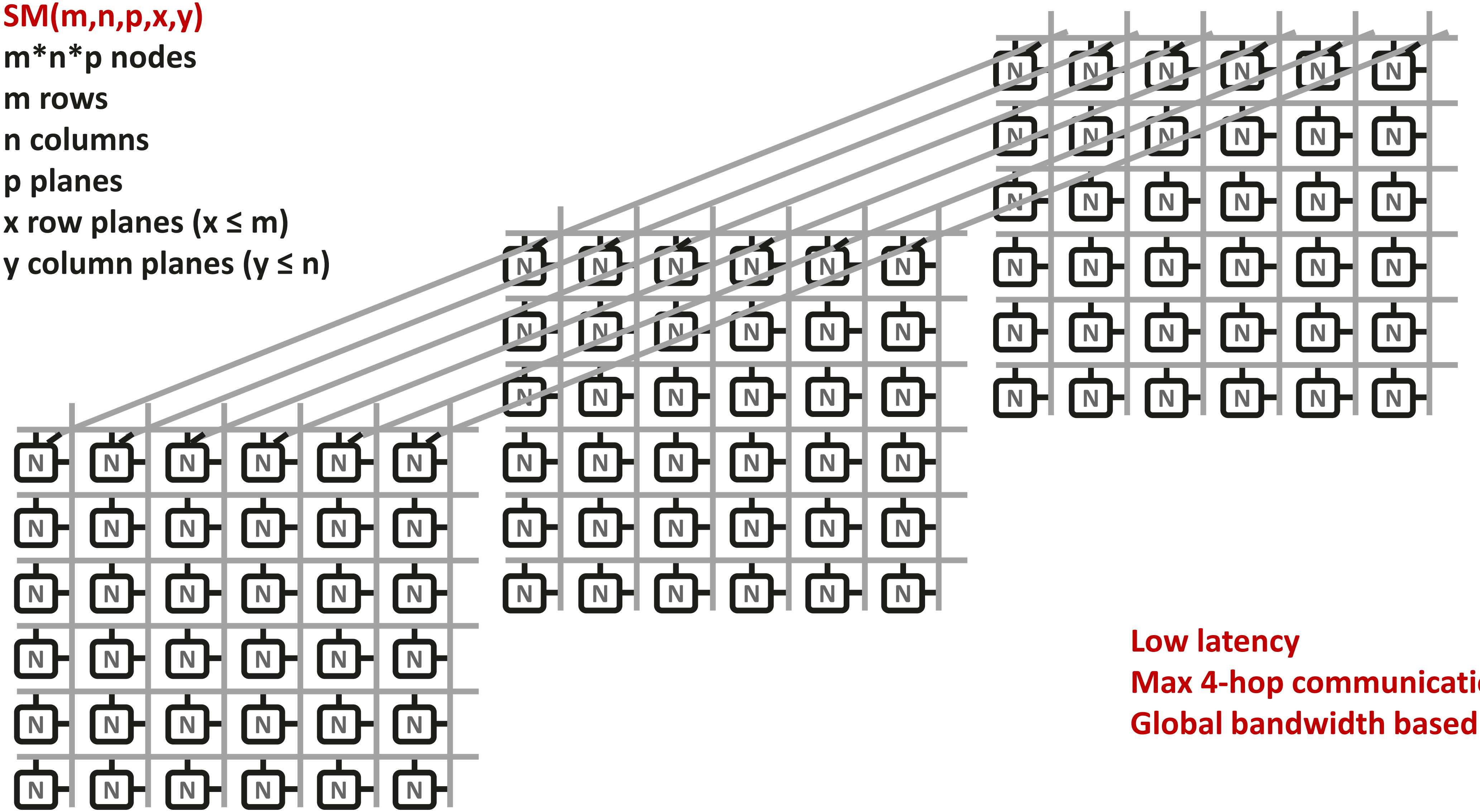


SM(6,6,3,1,0)



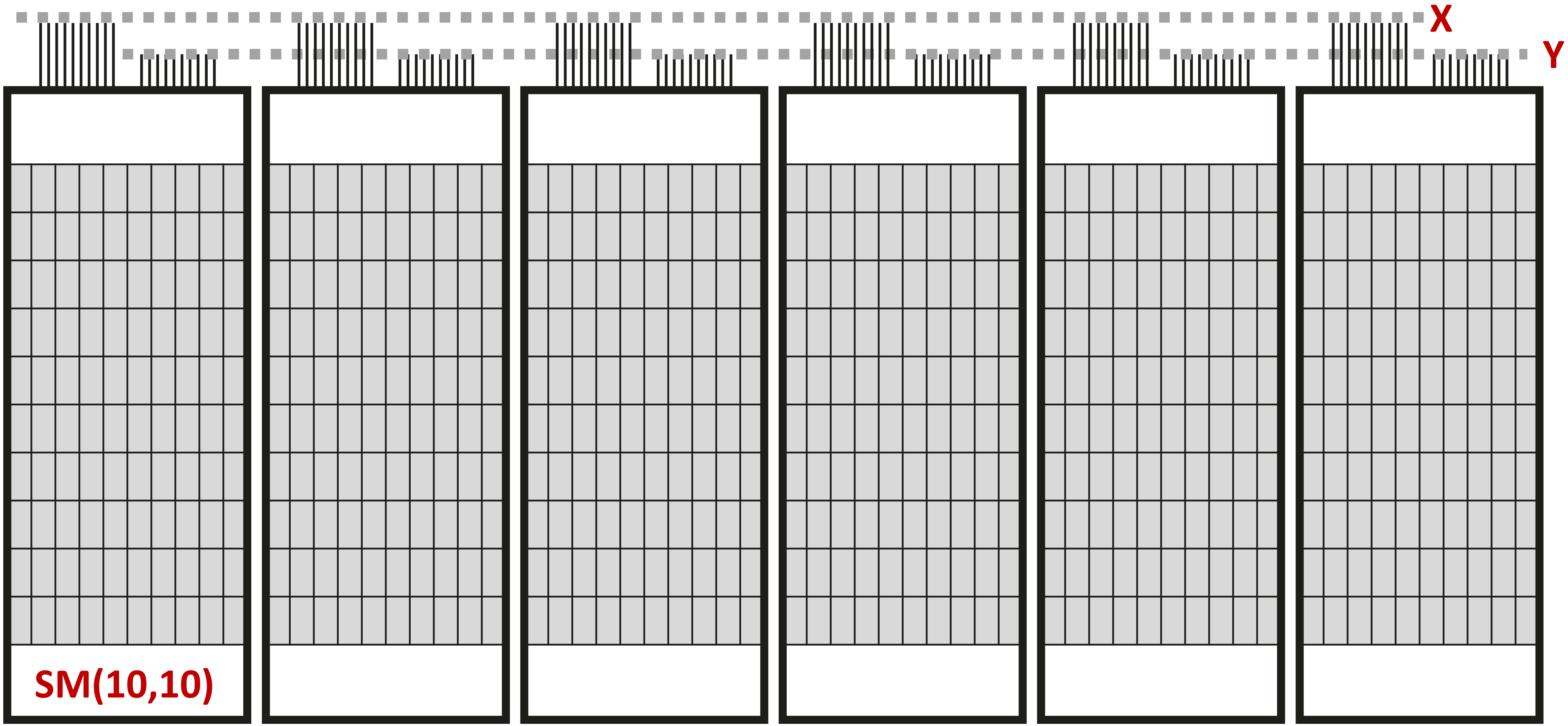
3D Supermesh SM(6,6,3,1,0)

SM(m,n,p,x,y)
m*n*p nodes
m rows
n columns
p planes
x row planes (x ≤ m)
y column planes (y ≤ n)

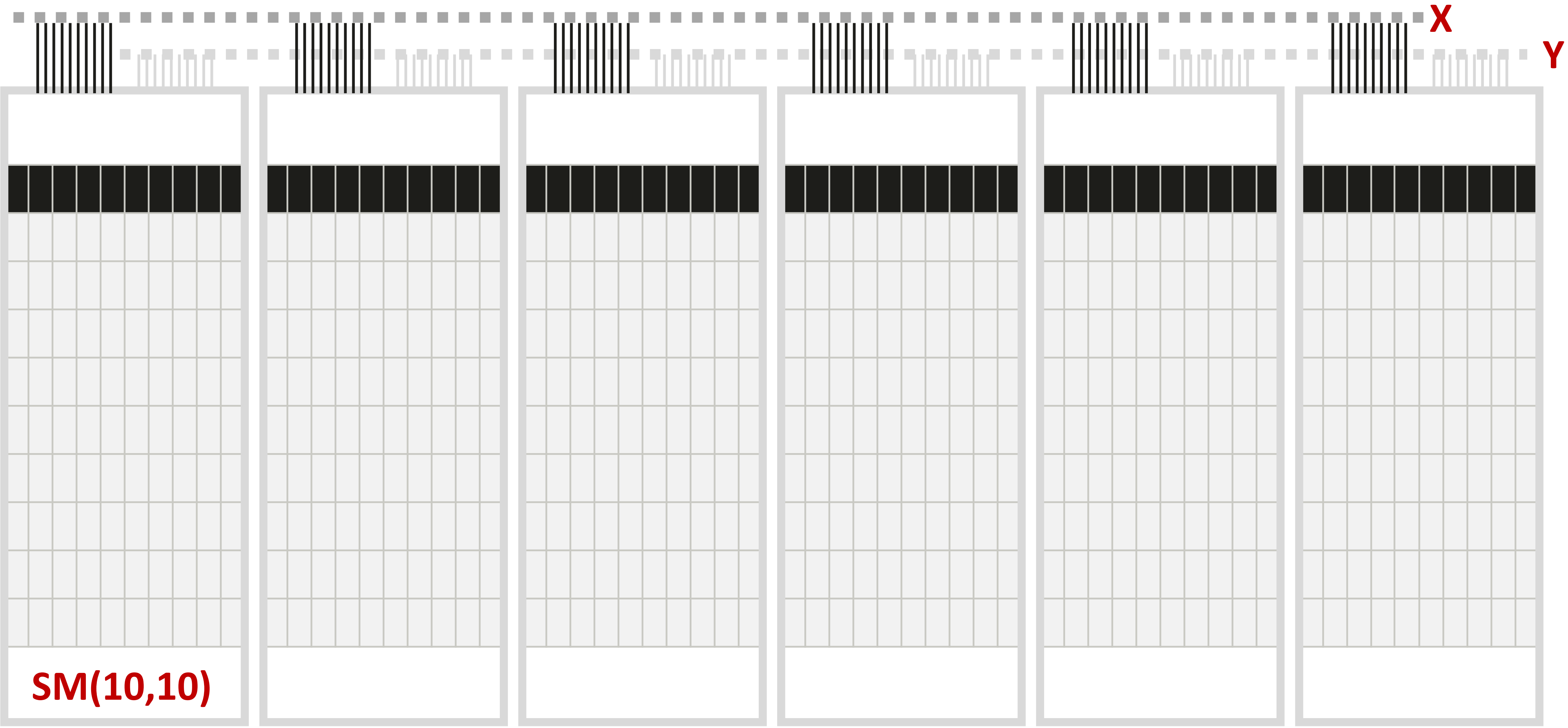


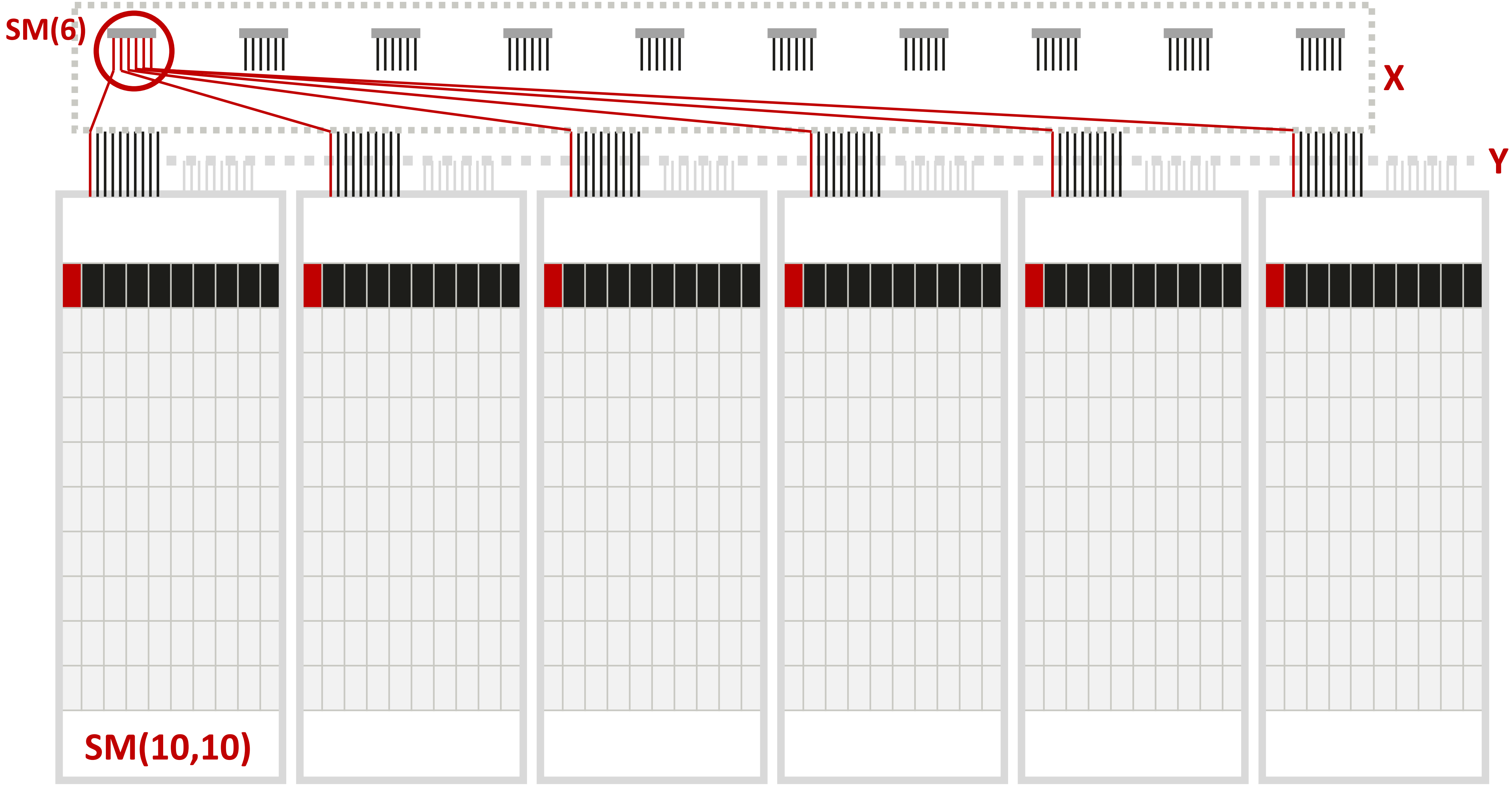
Low latency
Max 4-hop communication
Global bandwidth based on choice of x,y

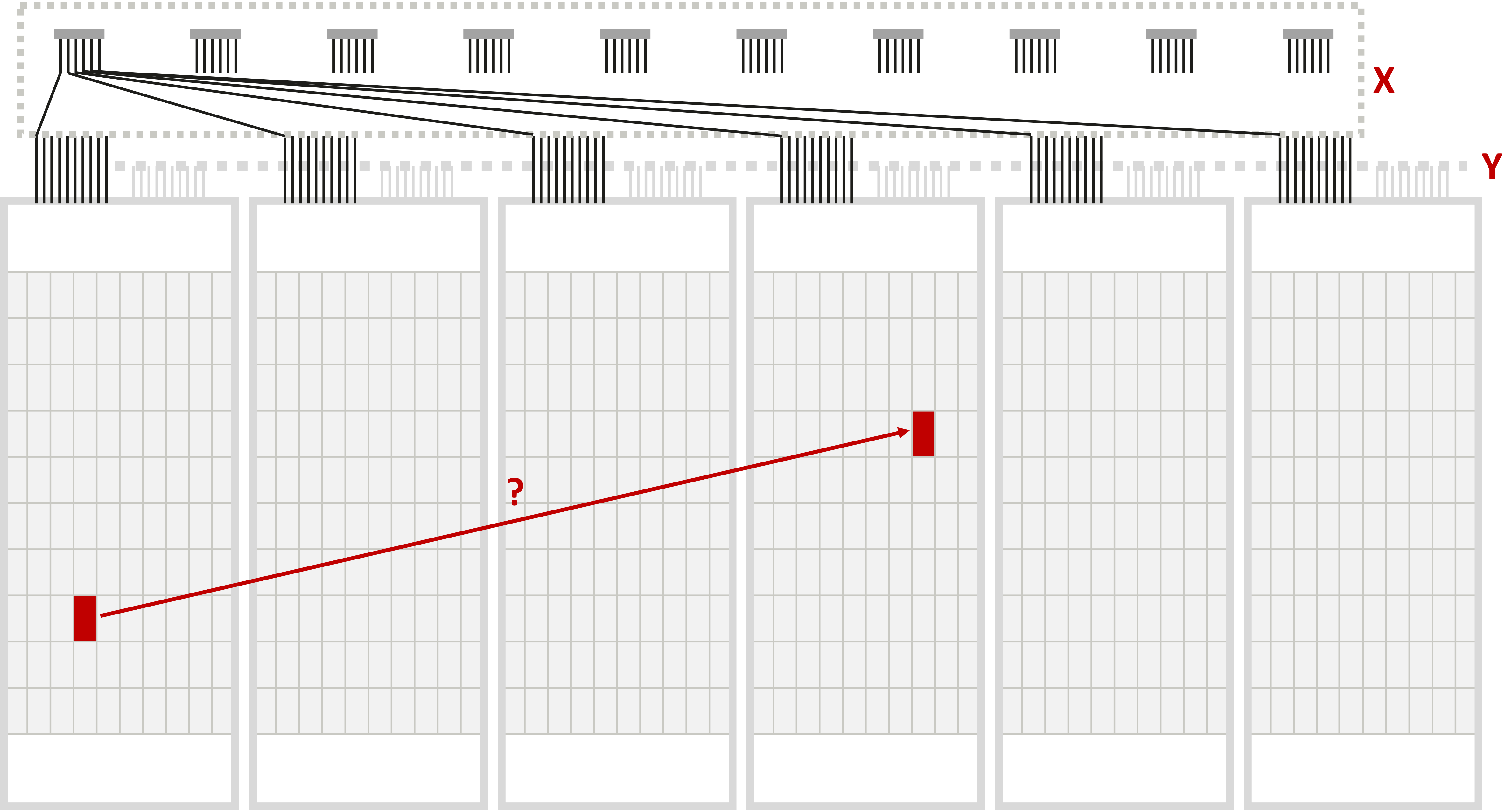
600-node Supermesh SM(10,10,6,1,1) in 6 racks

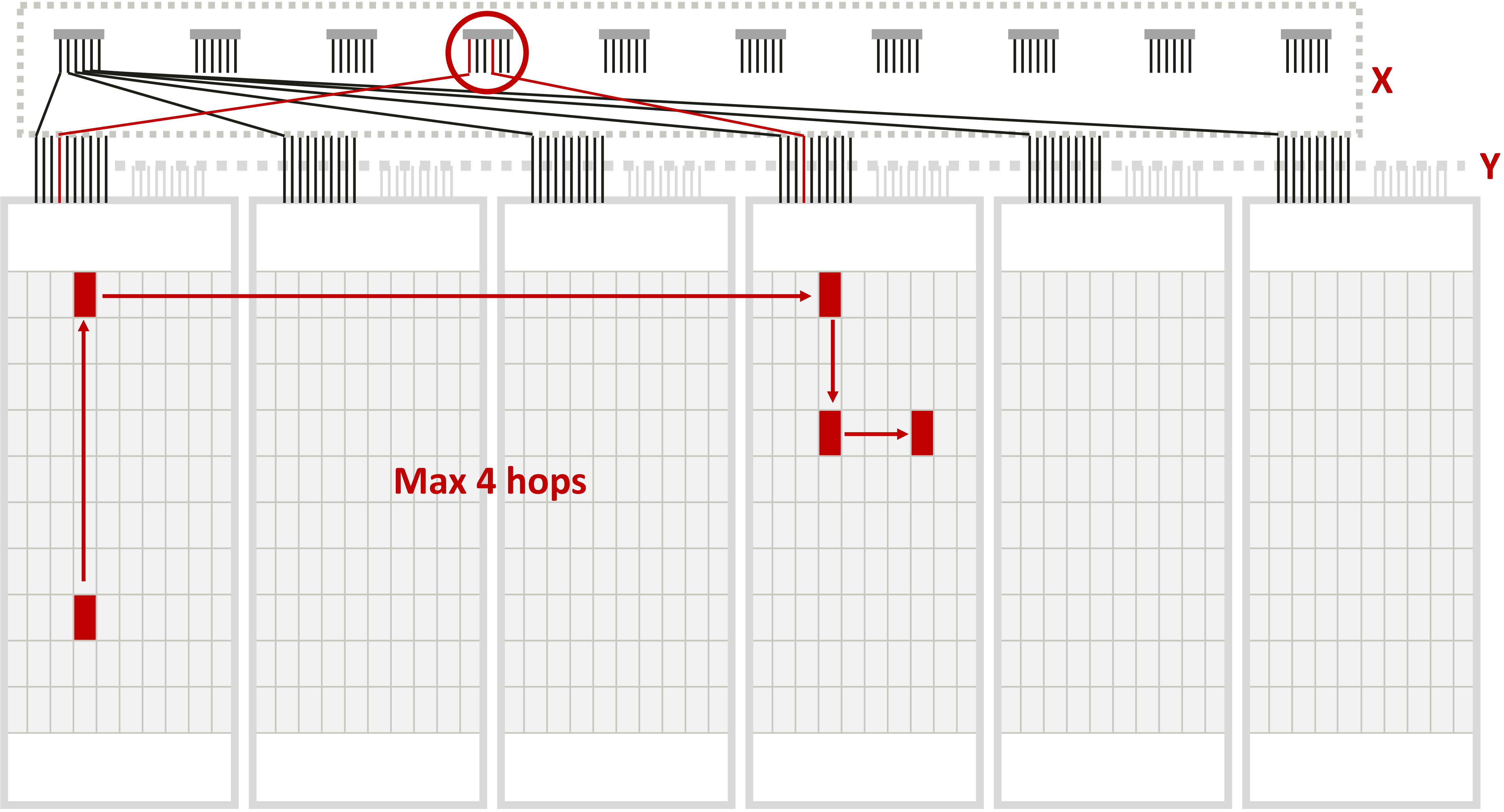


600-node SM(10,10,6,1,1) in 6 racks




SM(6)
X
Y
SM(10,10)


X
Y
?


X
Y
Max 4 hops

Global Bandwidth

Let w(x) = $\lfloor x/2 \rfloor * \lceil x/2 \rceil$ = $\lfloor x^2/4 \rfloor$.

For 3D SM, when m,n>1

GBW(SM(m,n,p,x,y)) = min { w(m)np , w(n)mp , w(p)(xn+ym) }

GBW(SM(1,n,p,x,y)) = min { w(n)p , w(p)(xn+y) }

GBW(SM(m,1,p,x,y)) = min { w(m)p , w(p)(x+ym) }

For 1D, since SM(m) = SM(m,1,1,0,0), GBW(SM(m)) = w(m) [Bisection Width]

For 2D, since SM(m,n) = SM(m,n,1,0,0), GBW(SM(m)) = min { w(m)n , w(n)m } [Bisection Width]

Virtual Supermeshes

Dedicated compute, memory and network

Fast : No link sharing (node-disjoint and edge-disjoint)

Flexible: No “contiguous nodes” requirements

Predictable: No app congestion

21 VSM domains on a single shared SM(10,10)

11	1	1	1	5	5	11	13	12	12
17	1	1	1	5	5	2	2	2	2
17	1	1	1	5	5	2	2	2	2
10	1	1	1	10	10	10	14	12	12
9	9	9	18	4	4	4	4	4	4
3	3	19	3	15	15	2	2	2	2
3	3	20	3	16	16	2	2	2	2
0	0	0	0	0	0	0	0	0	0
9	9	9	8	4	4	4	4	4	4
6	6	6	8	6	6	6	7	6	6

App	#Nodes
0	10
1	12
2	16
3	6
4	12
5	6
6	8
7	1
8	2
9	6
10	4
11	2
12	4
13	1
14	1
15	2
16	2
17	2
18	1
19	1
20	1

21 VSM domains on a single SM(10,10)

11	1	1	1	5	5	11	13	12	12
17	1	1	1	5	5	2	2	2	2
17	1	1	1	5	5	2	2	2	2
10	1	1	1	10	10	10	14	12	12
9	9	9	18	4	4	4	4	4	4
3	3	19	3	15	15	2	2	2	2
3	3	20	3	16	16	2	2	2	2
0	0	0	0	0	0	0	0	0	0
9	9	9	8	4	4	4	4	4	4
6	6	6	8	6	6	6	7	6	6

App	#Nodes
0	10
1	12
2	16
3	6
4	12
5	6
6	8
7	1
8	2
9	6
10	4
11	2
12	4
13	1
14	1
15	2
16	2
17	2
18	1
19	1
20	1

2 Nodes

1 Hop

21 VSM domains on a single SM(10,10)

11	1	1	1	5	5	11	13	12	12
17	1	1	1	5	5	2	2	2	2
17	1	1	1	5	5	2	2	2	2
10	1	1	1	10	10	10	14	12	12
9	9	9	18	4	4	4	4	4	4
3	3	19	3	15	15	2	2	2	2
3	3	20	3	16	16	2	2	2	2
0	0	0	0	0	0	0	0	0	0
9	9	9	8	4	4	4	4	4	4
6	6	6	8	6	6	6	7	6	6

App	#Nodes
0	10
1	12
2	16
3	6
4	12
5	6
6	8
7	1
8	2
9	6
10	4
11	2
12	4
13	1
14	1
15	2
16	2
17	2
18	1
19	1
20	1

8 Nodes

1 Hop

21 VSM domains on a single SM(10,10)

11	1	1	1	5	5	11	13	12	12
17	1	1	1	5	5	2	2	2	2
17	1	1	1	5	5	2	2	2	2
10	1	1	1	10	10	10	14	12	12
9	9	9	18	4	4	4	4	4	4
3	3	19	3	15	15	2	2	2	2
3	3	20	3	16	16	2	2	2	2
0	0	0	0	0	0	0	0	0	0
9	9	9	8	4	4	4	4	4	4
6	6	6	8	6	6	6	7	6	6

App	#Nodes
0	10
1	12
2	16
3	6
4	12
5	6
6	8
7	1
8	2
9	6
10	4
11	2
12	4
13	1
14	1
15	2
16	2
17	2
18	1
19	1
20	1

16 Nodes

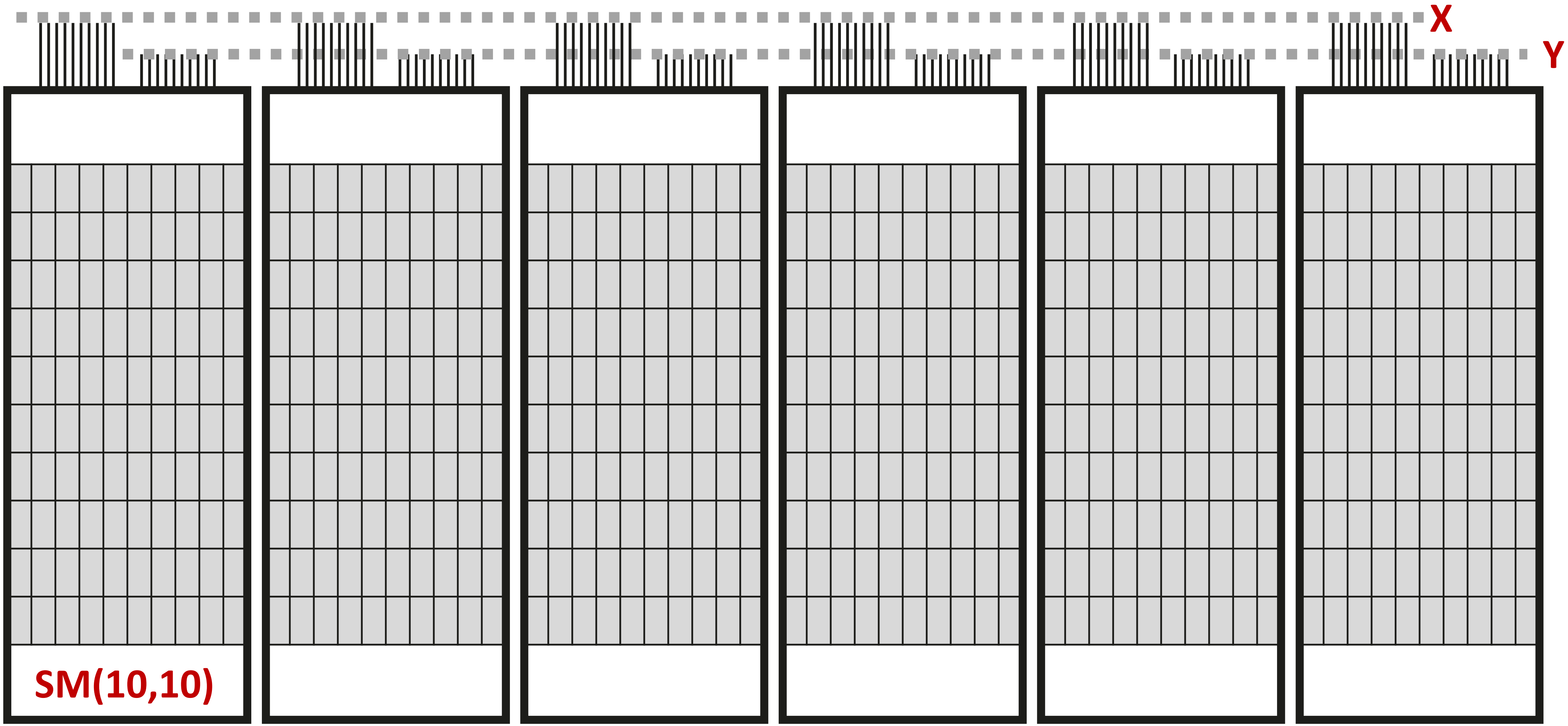
Max 2 Hops

21 VSM domains on a single SM(10,10)

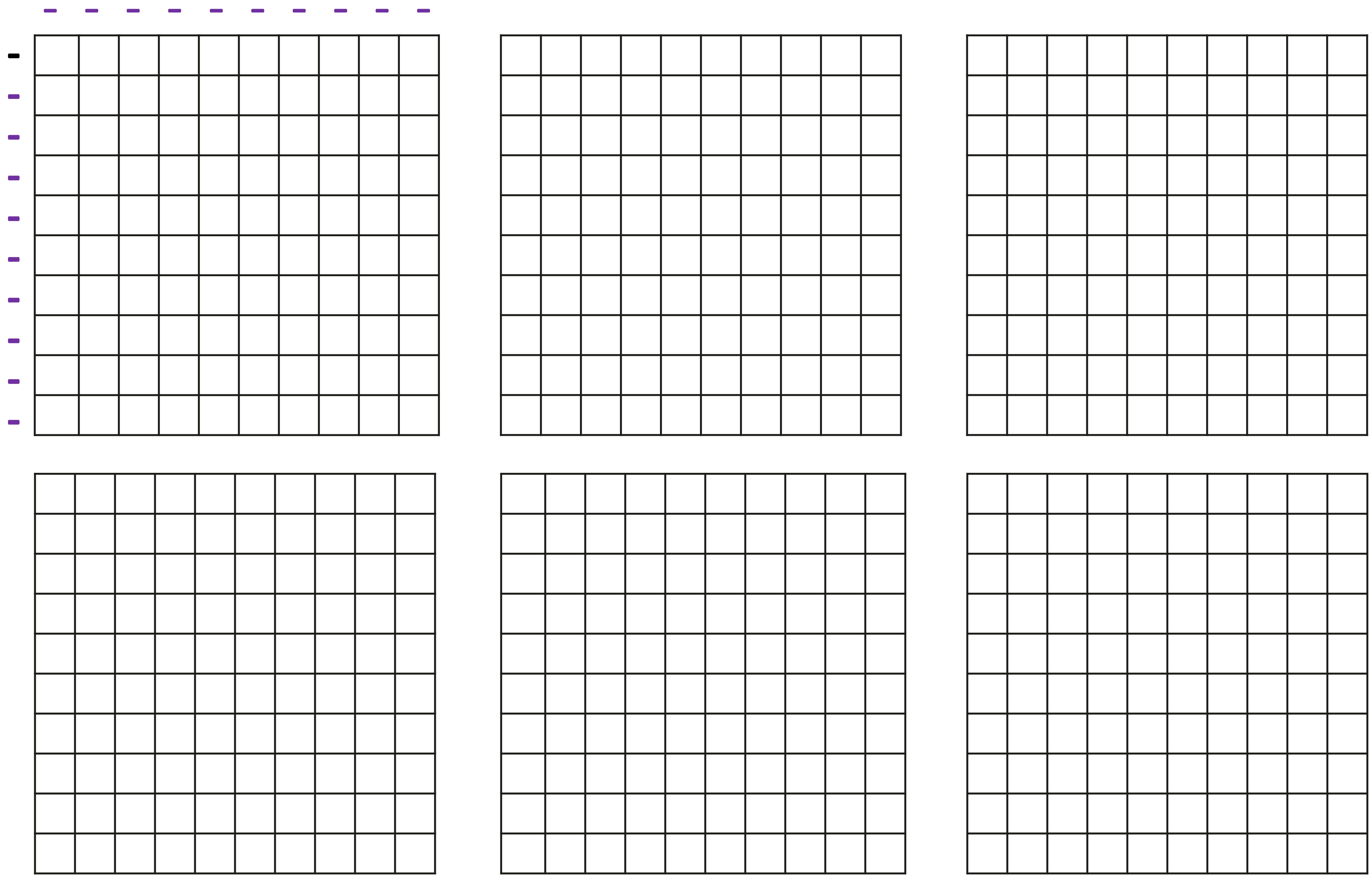
11	1	1	1	5	5	11	13	12	12
17	1	1	1	5	5	2	2	2	2
17	1	1	1	5	5	2	2	2	2
10	1	1	1	10	10	10	14	12	12
9	9	9	18	4	4	4	4	4	4
3	3	19	3	15	15	2	2	2	2
3	3	20	3	16	16	2	2	2	2
0	0	0	0	0	0	0	0	0	0
9	9	9	8	4	4	4	4	4	4
6	6	6	8	6	6	6	7	6	6

App	#Nodes
0	10
1	12
2	16
3	6
4	12
5	6
6	8
7	1
8	2
9	6
10	4
11	2
12	4
13	1
14	1
15	2
16	2
17	2
18	1
19	1
20	1

600-node Supermesh SM(10,10,6,1,1) in 6 racks



Mapping 8 VSMs onto a 600-node SM(10,10,6,1,1)



App	#Nodes
1	108
5	120
11	18
17	48
31	40
37	100
42	96
47	7

Mapping 8 VSMs onto a 600-node SM(10,10,6,1,1)

	5	1	1	1	37	37	5	5	5	31
-	5	1	1	1	37	37	5	5	5	31
-	5	1	1	1	37	37	5	5	5	31
17	17	1	1	1	37	37	17	17	17	31
17	17	1	1	1	37	37	17	17	17	31
42	42	42	42	42	37	37	42	42	42	42
-	5	1	1	1	37	37	5	5	5	31
-	5				37	37	5	5	5	31
42	42	42	42	42	37	37	42	42	42	42
-		1	1	1	37	37				31
-	5				37	37	5	5	5	31

5	1	1	1	11	11	5	5	5	31
5	1	1	1	11	11	5	5	5	31
17	1	1	1	11	11	17	17	17	31
17	1	1	1	11	11	17	17	17	31
42	42	42	42	11	11	42	42	42	42
5	1	1	1	11	11	5	5	5	31
5				11	11	5	5	5	31
42	42	42	42	11	11	42	42	42	42
	1	1	1	11	11				31
5						5	5	5	31

5	1	1	1	37	37	5	5	5	31
5	1	1	1	37	37	5	5	5	31
17	1	1	1	37	37	17	17	17	31
17	1	1	1	37	37	17	17	17	31
42	42	42	42	37	37	42	42	42	42
5	1	1	1	37	37	5	5	5	31
5				37	37	5	5	5	31
42	42	42	42	37	37	42	42	42	42
	1	1	1	37	37				31
5				37	37	5	5	5	31

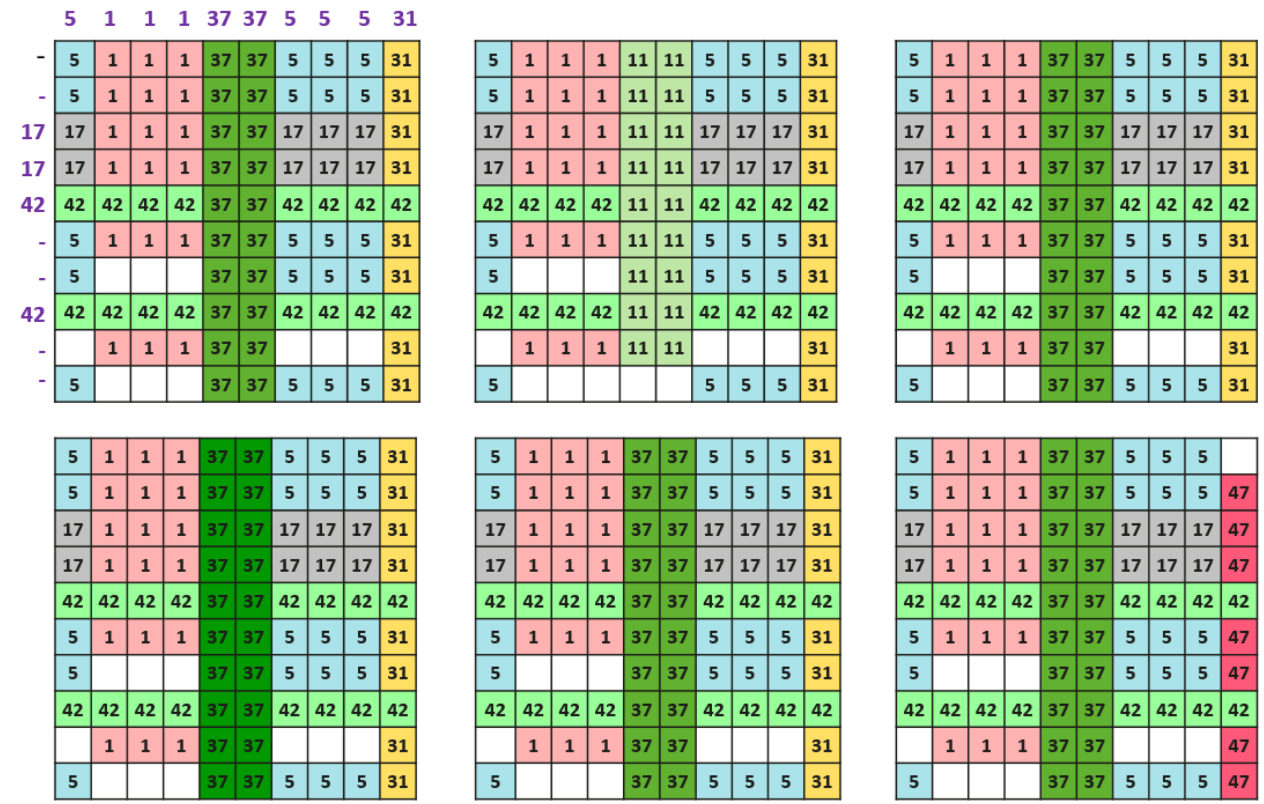
App	#Nodes
1	108
5	120
11	18
17	48
31	40
37	100
42	96
47	7

5	1	1	1	37	37	5	5	5	31
5	1	1	1	37	37	5	5	5	31
17	1	1	1	37	37	17	17	17	31
17	1	1	1	37	37	17	17	17	31
42	42	42	42	37	37	42	42	42	42
5	1	1	1	37	37	5	5	5	31
5				37	37	5	5	5	31
42	42	42	42	37	37	42	42	42	42
	1	1	1	37	37				31
5				37	37	5	5	5	31

5	1	1	1	37	37	5	5	5	31
5	1	1	1	37	37	5	5	5	31
17	1	1	1	37	37	17	17	17	31
17	1	1	1	37	37	17	17	17	31
42	42	42	42	37	37	42	42	42	42
5	1	1	1	37	37	5	5	5	31
5				37	37	5	5	5	31
42	42	42	42	37	37	42	42	42	42
	1	1	1	37	37				31
5				37	37	5	5	5	31

5	1	1	1	37	37	5	5	5	
5	1	1	1	37	37	5	5	5	47
17	1	1	1	37	37	17	17	17	47
17	1	1	1	37	37	17	17	17	47
42	42	42	42	37	37	42	42	42	42
5	1	1	1	37	37	5	5	5	47
5				37	37	5	5	5	47
42	42	42	42	37	37	42	42	42	42
	1	1	1	37	37				47
5				37	37	5	5	5	47

Mapping 8 VSMs onto a 600-node SM(10,10,6,1,1)



App	VSM	#Nodes	Diameter	GBW
1	$VSM(6, 3, 6, 1, 0)$	108	4	27
5	$VSM(5, 4, 6, 1, 0)$	120	4	36
11	$VSM(9, 2)$	18	2	9
17	$VSM(2, 4, 6, 0, 1)$	48	4	18
31	$VSM(8, 1, 5, 1, 0)$	40	3	6
37	$VSM(10, 2, 5, 1, 0)$	100	4	12
42	$VSM(2, 8, 6, 0, 1)$	96	4	18
47	$VSM(7)$	7	1	12

Another Mapping...

	5	5	5	5	5	5	5	5	5	5
1	1	1	1	1	1	1	1	1	1	47
1	1	1	1	1	1	1	1	1	1	47
1	1	1	1	1	1	1	1	1	1	47
1	1	1	1	1	1	1	1	1	1	47
1	1	1	1	1	1	1	1	1	1	47
1	1	1	1	1	1	1	1	1	1	47
42	42	42	42	42	42	42	42	42		47
42	42	42	42	42	42	42	42	42		
42	42	42	42	42	42	42	42	42		
42	42	42	42	42	42	42	42	42		

1	1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	1	
42	42	42	42	42	42	42	42		
42	42	42	42	42	42	42	42		
42	42	42	42	42	42	42	42		
42	42	42	42	42	42	42	42		

5	5	5	5	5	5	5	5	5	5
5	5	5	5	5	5	5	5	5	5
5	5	5	5	5	5	5	5	5	5
5	5	5	5	5	5	5	5	5	5
5	5	5	5	5	5	5	5	5	5
5	5	5	5	5	5	5	5	5	5
42	42	42	42	42	42	42	42		
42	42	42	42	42	42	42	42		
42	42	42	42	42	42	42	42		
42	42	42	42	42	42	42	42		

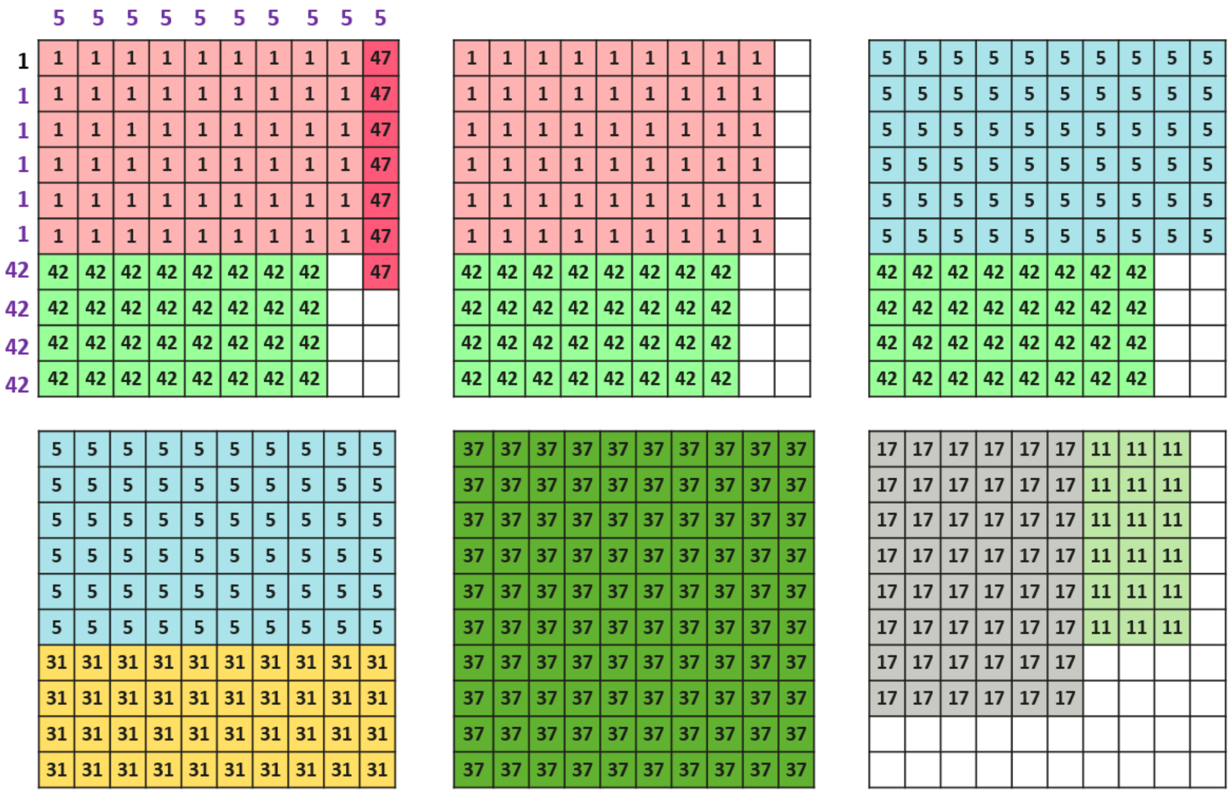
5	5	5	5	5	5	5	5	5	5
5	5	5	5	5	5	5	5	5	5
5	5	5	5	5	5	5	5	5	5
5	5	5	5	5	5	5	5	5	5
5	5	5	5	5	5	5	5	5	5
5	5	5	5	5	5	5	5	5	5
31	31	31	31	31	31	31	31	31	31
31	31	31	31	31	31	31	31	31	31
31	31	31	31	31	31	31	31	31	31
31	31	31	31	31	31	31	31	31	31

37	37	37	37	37	37	37	37	37	37
37	37	37	37	37	37	37	37	37	37
37	37	37	37	37	37	37	37	37	37
37	37	37	37	37	37	37	37	37	37
37	37	37	37	37	37	37	37	37	37
37	37	37	37	37	37	37	37	37	37
37	37	37	37	37	37	37	37	37	37
37	37	37	37	37	37	37	37	37	37
37	37	37	37	37	37	37	37	37	37
37	37	37	37	37	37	37	37	37	37

17	17	17	17	17	17	11	11	11	
17	17	17	17	17	17	11	11	11	
17	17	17	17	17	17	11	11	11	
17	17	17	17	17	17	11	11	11	
17	17	17	17	17	17	11	11	11	
17	17	17	17	17	17	11	11	11	
17	17	17	17	17	17				
17	17	17	17	17	17				

App	#Nodes
1	108
5	120
11	18
17	48
31	40
37	100
42	96
47	7

Another Mapping…



App	VSM	#Nodes	Diameter	GBW
1	$VSM(6, 9, 2, 0, 1)$	108	4	6
5	$VSM(6, 10, 2, 1, 0)$	120	4	10
11	$VSM(6, 3)$	18	2	12
17	$VSM(8, 6)$	48	2	72
31	$VSM(4, 10)$	40	2	40
37	$VSM(10, 10)$	100	2	250
42	$VSM(4, 8, 3, 0, 1)$	96	4	8
47	$VSM(7)$	7	1	12

A Third Mapping...

	5	5	5	5	5	5	5	5	5	5
-	5	5	5	5	5	5	5	5	5	5
-	5	5	5	5	5	5	5	5	5	5
-	5	5	5	5	5	5	5	5	5	5
1	1	1	1	1	1	1	1	1	1	47
1	1	1	1	1	1	1	1	1	1	47
1	1	1	1	1	1	1	1	1	1	47
42	42	42	42	42	42	42				47
42	42	42	42	42	42	42				47
42	42	42	42	42	42	42				47
42	42	42	42	42	42	42				47

5	5	5	5	5	5	5	5	5	5
5	5	5	5	5	5	5	5	5	5
5	5	5	5	5	5	5	5	5	5
1	1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	1	
42	42	42	42	42	42				
42	42	42	42	42	42				
42	42	42	42	42	42				
42	42	42	42	42	42				

5	5	5	5	5	5	5	5	5	5
5	5	5	5	5	5	5	5	5	5
5	5	5	5	5	5	5	5	5	5
1	1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	1	
42	42	42	42	42	42				
42	42	42	42	42	42				
42	42	42	42	42	42				
42	42	42	42	42	42				

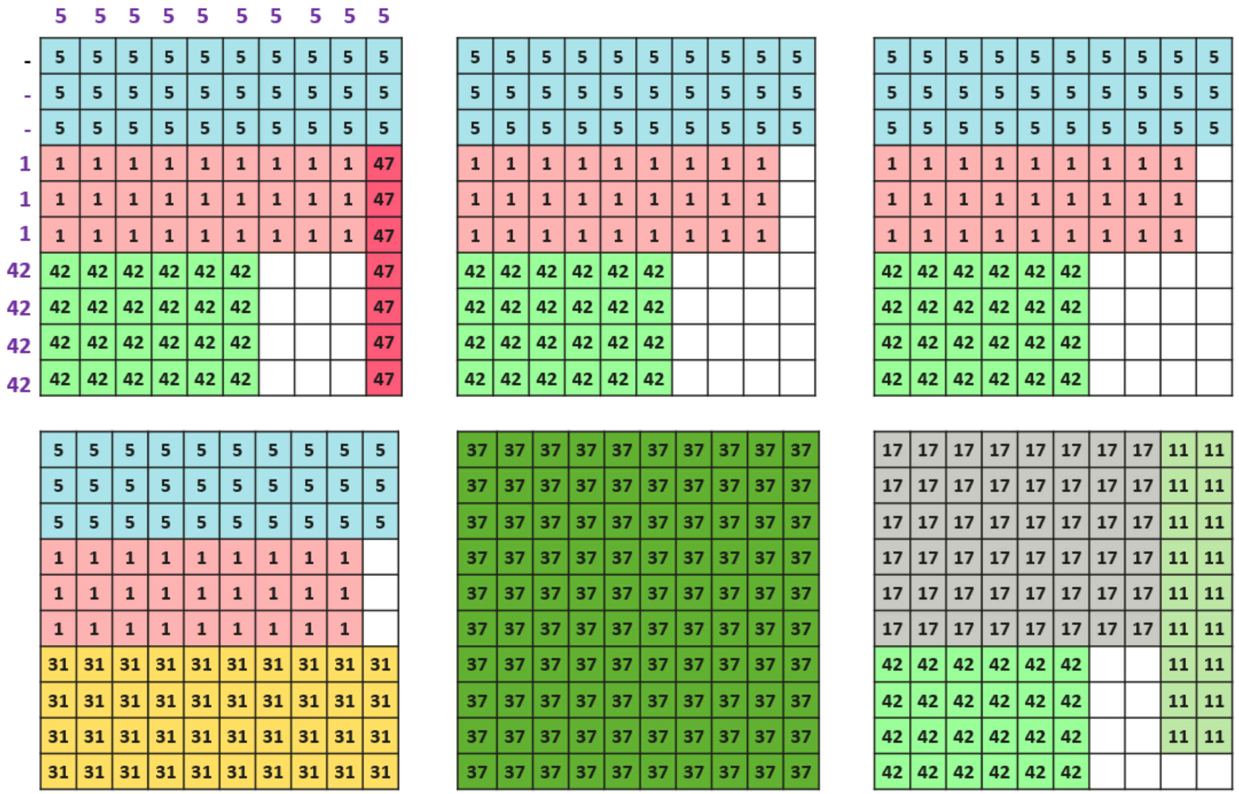
5	5	5	5	5	5	5	5	5	5
5	5	5	5	5	5	5	5	5	5
5	5	5	5	5	5	5	5	5	5
1	1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	1	
31	31	31	31	31	31	31	31	31	31
31	31	31	31	31	31	31	31	31	31
31	31	31	31	31	31	31	31	31	31
31	31	31	31	31	31	31	31	31	31

37	37	37	37	37	37	37	37	37	37
37	37	37	37	37	37	37	37	37	37
37	37	37	37	37	37	37	37	37	37
37	37	37	37	37	37	37	37	37	37
37	37	37	37	37	37	37	37	37	37
37	37	37	37	37	37	37	37	37	37
37	37	37	37	37	37	37	37	37	37
37	37	37	37	37	37	37	37	37	37
37	37	37	37	37	37	37	37	37	37
37	37	37	37	37	37	37	37	37	37

17	17	17	17	17	17	17	17	11	11
17	17	17	17	17	17	17	17	11	11
17	17	17	17	17	17	17	17	11	11
17	17	17	17	17	17	17	17	11	11
17	17	17	17	17	17	17	17	11	11
17	17	17	17	17	17	17	17	11	11
42	42	42	42	42	42			11	11
42	42	42	42	42	42			11	11
42	42	42	42	42	42			11	11
42	42	42	42	42	42				

App	#Nodes
1	108
5	120
11	18
17	48
31	40
37	100
42	96
47	7

A Third Mapping...



App	VSM	#Nodes	Diameter	GBW
1	$VSM(3, 9, 4, 0, 1)$	108	4	12
5	$VSM(3, 10, 4, 1, 0)$	120	4	40
11	$VSM(9, 2)$	18	2	9
17	$VSM(6, 8)$	48	2	72
31	$VSM(4, 10)$	40	2	40
37	$VSM(10, 10)$	100	2	250
42	$VSM(4, 6, 4, 0, 1)$	96	4	16
47	$VSM(7)$	7	1	12

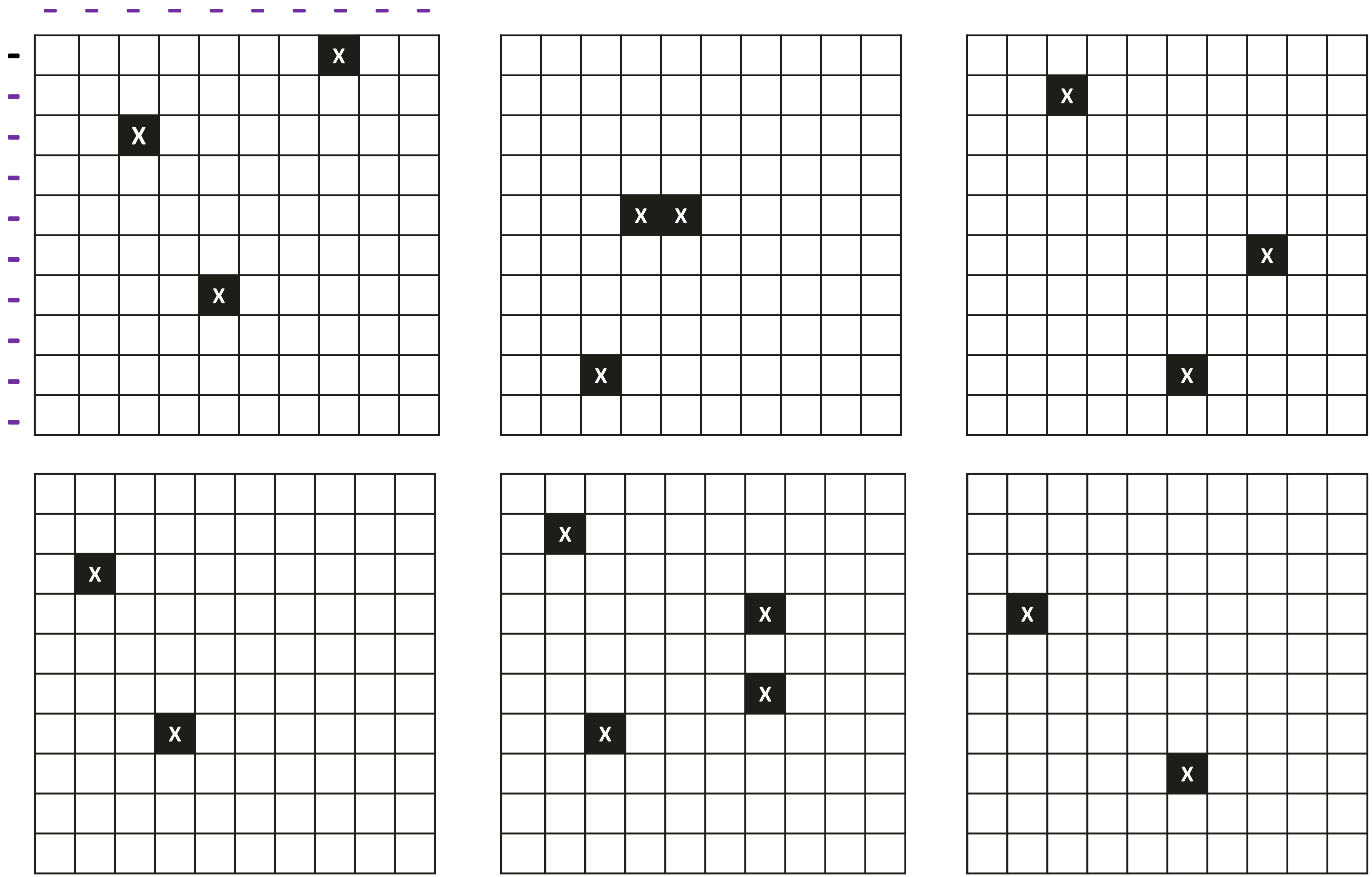
Tradeoffs…

App	VSM	#Nodes	Diameter	GBW
1	$VSM(6,3,6,1,0)$	108	4	27
5	$VSM(5,4,6,1,0)$	120	4	36
11	$VSM(9,2)$	18	2	9
17	$VSM(2,4,6,0,1)$	48	4	18
31	$VSM(8,1,5,1,0)$	40	3	6
37	$VSM(10,2,5,1,0)$	100	4	12
42	$VSM(2,8,6,0,1)$	96	4	18
47	$VSM(7)$	7	1	12

App	VSM	#Nodes	Diameter	GBW
1	$VSM(6,9,2,0,1)$	108	4	6
5	$VSM(6,10,2,1,0)$	120	4	10
11	$VSM(6,3)$	18	2	12
17	$VSM(8,6)$	48	2	72
31	$VSM(4,10)$	40	2	40
37	$VSM(10,10)$	100	2	250
42	$VSM(4,8,3,0,1)$	96	4	8
47	$VSM(7)$	7	1	12

App	VSM	#Nodes	Diameter	GBW
1	$VSM(3,9,4,0,1)$	108	4	12
5	$VSM(3,10,4,1,0)$	120	4	40
11	$VSM(9,2)$	18	2	9
17	$VSM(6,8)$	48	2	72
31	$VSM(4,10)$	40	2	40
37	$VSM(10,10)$	100	2	250
42	$VSM(4,6,4,0,1)$	96	4	16
47	$VSM(7)$	7	1	12

Mapping With Faults



Mapping 17 VSMs onto a 600-node SM(10,10,6,1,1) with Faults

	25	25	32	25	39	8	8	-	8	8
22	25	25	32	25	39	8	8	X	8	8
22	25	25		25	39	8	8	91	8	8
22	25	25	X	25	39	8	8	91	8	8
22	25	25	32	25	39	8	8	91	8	8
22	25	25	32	25		8	8	91	8	8
22	25	25	32	25	39	8	8	91	8	8
85	25	25	32	25	X	8	8	91	8	8
12	12	12	12	12	12	8	8	91	8	8
12	12	12	12	12	12	8	8	91	8	8
12	12	12	12	12	12	8	8	91	8	8

44	44	32		39	8	8	44	8	8
44	44			39	8	8	44	8	8
44	44			39	8	8	44	8	8
44	44	32		39	8	8	44	8	8
44	44	32	X	X	8	8	44	8	8
44	44	32		39	8	8	44	8	8
44	44	32	63	63	8	8	44	8	8
44	44		63	63	8	8	44	8	8
44	44	X	63	63	8	8	44	8	8
44	44		63	63	8	8	44	8	8

25	25	32	25	39	82	82	82	68	68
25	25	X	25	39	82	82	82	68	68
25	25		25	39	82	82	82	68	68
25	25	32	25	39	82	82	82	68	68
25	25	32	25		82	82	82	68	68
25	25	32	25	39			X	68	68
25	25	32	25		82	82	82	68	68
12	12	12	12	12	82	82	82	68	68
12	12	12	12	12	X			68	68
12	12	12	12	12	82	82	82	68	68

22		32	22	22	8	8		8	8
22			22	22	8	8		8	8
22	X		22	22	8	8		8	8
22		32	22	22	8	8		8	8
22		32	22	22	8	8		8	8
22		32	22	22	8	8		8	8
85	85	32	X	85	8	8		8	8
12	12	12	12	12	8	8		8	8
12	12	12	12	12	8	8		8	8
12	12	12	12	12	8	8		8	8

22			22	22	74	77	55	55	55
22	X		22	22	74	77	55	55	55
22			22	22	74	77	55	55	55
22			22	22	74	X	55	55	55
22			22	22	74	77	55	55	55
22			22	22	74	X	55	55	55
85	85	X		85	74	77	55	55	55
12	12	12	12	12	74	77	55	55	55
12	12	12	12	12	74	77	55	55	55
12	12	12	12	12	74	77	55	55	55

22		32	22	22	51	51	15	15	15
22			22	22	51	51	15	15	15
22			22	22	51	51	15	15	15
22	X	32	22	22	51	51	15	15	15
22		32	22	22	51	51	15	15	15
22		32	22	22	51	51	15	15	15
85	85	32		85	51	51	15	15	15
12	12	12	12	12	X		15	15	15
12	12	12	12	12			15	15	15
12	12	12	12	12			15	15	15

App	#Nodes
8	120
12	75
15	30
22	54
25	42
32	25
39	15
44	30
51	14
55	30
63	8
68	20
74	10
77	8
82	24
85	9
91	9

Mapping 17 VSMs onto a 600-node SM(10,10,6,1,1) with Faults

	25	25	32	25	39	8	8	-	8	8
22	25	25	32	25	39	8	8	X	8	8
22	25	25		25	39	8	8	91	8	8
22	25	25	X	25	39	8	8	91	8	8
22	25	25	32	25	39	8	8	91	8	8
22	25	25	32	25		8	8	91	8	8
22	25	25	32	25	39	8	8	91	8	8
85	25	25	32	25	X	8	8	91	8	8
12	12	12	12	12	12	8	8	91	8	8
12	12	12	12	12	12	8	8	91	8	8
12	12	12	12	12	12	8	8	91	8	8

44	44	32		39	8	8	44	8	8
44	44			39	8	8	44	8	8
44	44			39	8	8	44	8	8
44	44	32		39	8	8	44	8	8
44	44	32	X	X	8	8	44	8	8
44	44	32		39	8	8	44	8	8
44	44	32	63	63	8	8	44	8	8
44	44		63	63	8	8	44	8	8
44	44	X	63	63	8	8	44	8	8
44	44		63	63	8	8	44	8	8

25	25	32	25	39	82	82	82	68	68
25	25	X	25	39	82	82	82	68	68
25	25		25	39	82	82	82	68	68
25	25	32	25	39	82	82	82	68	68
25	25	32	25		82	82	82	68	68
25	25	32	25	39			X	68	68
25	25	32	25		82	82	82	68	68
12	12	12	12	12	82	82	82	68	68
12	12	12	12	12	X			68	68
12	12	12	12	12	82	82	82	68	68

22		32	22	22	8	8		8	8
22			22	22	8	8		8	8
22	X		22	22	8	8		8	8
22		32	22	22	8	8		8	8
22		32	22	22	8	8		8	8
22		32	22	22	8	8		8	8
85	85	32	X	85	8	8		8	8
12	12	12	12	12	8	8		8	8
12	12	12	12	12	8	8		8	8
12	12	12	12	12	8	8		8	8

22			22	22	74	77	55	55	55
22	X		22	22	74	77	55	55	55
22			22	22	74	77	55	55	55
22			22	22	74	X	55	55	55
22			22	22	74	77	55	55	55
22			22	22	74	X	55	55	55
85	85	X		85	74	77	55	55	55
12	12	12	12	12	74	77	55	55	55
12	12	12	12	12	74	77	55	55	55
12	12	12	12	12	74	77	55	55	55

22		32	22	22	51	51	15	15	15
22			22	22	51	51	15	15	15
22			22	22	51	51	15	15	15
22	X	32	22	22	51	51	15	15	15
22		32	22	22	51	51	15	15	15
22		32	22	22	51	51	15	15	15
85	85	32		85	51	51	15	15	15
12	12	12	12	12	X		15	15	15
12	12	12	12	12			15	15	15
12	12	12	12	12			15	15	15

App	VSM	#Nodes	Diameter	GBW
8	$VSM(10,4,3,1,0)$	120	4	8
12	$VSM(3,5,5,0,1)$	75	4	18
15	$VSM(10,3)$	30	2	20
22	$VSM(6,3,3,0,1)$	54	4	12
25	$VSM(7,3,2,1,0)$	42	4	3
32	$VSM(5,1,5,1,0)$	25	3	6
39	$VSM(5,1,3,1,0)$	15	3	2
44	$VSM(10,3)$	30	2	20
51	$VSM(7,2)$	14	2	7
55	$VSM(10,3)$	30	2	20
63	$VSM(4,2)$	8	2	4
68	$VSM(10,2)$	20	2	10
74	$VSM(10)$	10	1	25
77	$VSM(8)$	8	1	16
82	$VSM(8,3)$	24	2	16
85	$VSM(1,3,3,0,1)$	9	3	2
91	$VSM(9)$	9	1	20

Supermesh Communication Cost Model

Cost of single round of data exchange = <u>Maximum data volume sent on any one link</u>

Cost of whole data exchange = <u>Sum of costs of rounds</u>

Minimum “sync” time for any round, e.g. time to send a small 10KB vector across a link

Cost model encourages ultra-fast parallel data exchanges, with concurrent use of all links in multiple dimensions

Cost model discourages inefficient exchanges, with only small data volumes moved

Communication on a 1D Supermesh S(8)

	0	1	2	3	4	5	6	7	
0	·			8	9	7		7	31
1	7	·	7		8		6	7	35
2	7		·	7	7		1	7	29
3		7	9	·			7		23
4		5		3	·	7		7	22
5	5	1	7		7	·	2		22
6		7	1	1	7		·	7	23
7				7		6	5	·	18
	19	20	24	26	38	20	21	35	

Nodes: *0..7*

Maximum message size: $M = 9$

Maximum data sent: $MaxR = 35$

Maximum data received: $MaxC = 38$

Communication on a 1D Supermesh S(8)

	0	1	2	3	4	5	6	7	
0	·			8	9	7		7	31
1	7	·	7		8		6	7	35
2	7		·	7	7		1	7	29
3		7	9	·			7		23
4		5		3	·	7		7	22
5	5	1	7		7	·	2		22
6		7	1	1	7		·	7	23
7				7		6	5	·	18
	19	20	24	26	38	20	21	35	

If

Nodes: *0..7*

Maximum message size: $M=9$

Maximum avg data sent: $R = \lceil MaxR/8 \rceil = 5$

Maximum avg data received: $C = \lceil MaxC/8 \rceil = 5$

Direct and Indirect All-to-All on a 1D Supermesh S(m)

Direct (DA2A)

One round / single-hop

Send messages directly

Cost M

Best for small messages and for certain regular patterns

Indirect (IA2A)

Two rounds / 2-hops

Round1: Scatter a fraction $1/m$ of each outgoing message, as a set of sub-messages, across the m nodes. Cost R.

Round2: Forward sub-messages to destinations. Cost C.

Total Cost: $R+C$

Best for many/most irregular patterns

If $M \leq R+C$ then use DA2A else use IA2A

Communication on a 1D Supermesh S(8)

	0	1	2	3	4	5	6	7	
0	·			8	9	7		7	31
1	7	·	7		8		6	7	35
2	7		·	7	7		1	7	29
3		7	9	·			7		23
4		5		3	·	7		7	22
5	5	1	7		7	·	2		22
6		7	1	1	7		·	7	23
7				7		6	5	·	18
	19	20	24	26	38	20	21	35	

$M=9$

$R=\lceil 35/8 \rceil = 5$

$C=\lceil 38/8 \rceil = 5$

$M<R+C$ → DA2A

Regular 1-Relation (Permutation)

	0	1	2	3	4	5	6	7	
0	·			5					5
1	5	·							5
2			·			5			5
3			5	·					5
4					·		5		5
5					5	·			5
6							·	5	5
7		5						·	5
	5	5	5	5	5	5	5	5	

$M=5$

$R = \lceil 5/8 \rceil = 1$

$C = \lceil 5/8 \rceil = 1$

$M > R+C \rightarrow$ IA2A

Regular 2-Relation

	0	1	2	3	4	5	6	7	
0	·			5	5				10
1	5	·						5	10
2			·			5	5		10
3		5	5	·					10
4	5				·		5		10
5			5		5	·			10
6				5			·	5	10
7		5				5		·	10
	10	10	10	10	10	10	10	10	

$M=5$

$R=\lceil 10/8 \rceil = 2$

$C=\lceil 10/8 \rceil = 2$

$M > R+C \rightarrow$ IA2A

Regular 4-Relation

	0	1	2	3	4	5	6	7	
0	·		8	8			8	8	32
1	8	·		8		8		8	32
2		8	·		8	8	8		32
3	8	8		·	8			8	32
4		8	8	8	·		8		32
5	8		8		8	·	8		32
6		8	8			8	·	8	32
7	8			8	8	8		·	32
	32	32	32	32	32	32	32	32	

$M=8$

$R=\lceil 32/8 \rceil = 4$

$C=\lceil 32/8 \rceil = 4$

$M=R+C \rightarrow$ DA2A

Regular 7-Relation (All-to-All)

	0	1	2	3	4	5	6	7	
0	·	8	8	8	8	8	8	8	56
1	8	·	8	8	8	8	8	8	56
2	8	8	·	8	8	8	8	8	56
3	8	8	8	·	8	8	8	8	56
4	8	8	8	8	·	8	8	8	56
5	8	8	8	8	8	·	8	8	56
6	8	8	8	8	8	8	·	8	56
7	8	8	8	8	8	8	8	·	56
	56	56	56	56	56	56	56	56	

$M=8$

$R = \lceil 56/8 \rceil = 7$

$C = \lceil 56/8 \rceil = 7$

$M < R+C \rightarrow$ DA2A

Copy

	0	1	2	3	4	5	6	7	
0	·								
1		·							
2			·						
3				·					
4					·				
5						·			
6				8			·		8
7								·	
				8					

$M = 8$

$R = \lceil 8/8 \rceil = 1$

$C = \lceil 8/8 \rceil = 1$

$M > R + C$ → IA2A

Copy a single vector by Scatter+Gather.
NOTE. This kind of ultra-sparse communication is inefficient and should be avoided by algorithm and software designers.

Scatter

	0	1	2	3	4	5	6	7	
0	·								
1		·							
2			·						
3				·					
4					·				
5						·			
6	8	8	8	8	8	8	·	8	56
7								·	
	8	8	8	8	8	8		8	

$M=8$

$R=\lceil 56/8\rceil=7$

$C=\lceil 8/8\rceil=1$

$M=R+C$ → DA2A

Gather

	0	1	2	3	4	5	6	7	
0	·		8						8
1		·	8						8
2			·						
3			8	·					8
4			8		·				8
5			8			·			8
6			8				·		8
7			8					·	8
			56						

$M=8$

$R = \lceil 8/8 \rceil = 1$

$C = \lceil 56/8 \rceil = 7$

$M = R + C$ → DA2A

Communication Costs for m-node 1D Supermesh S(m)

Communication	Cost
h-Relation, $h < m/2$	$2V/m$
h-Relation, $h \geq m/2$	V/h
All-to-All	V/m
Copy	$2V/m$
Scatter	V/m
Gather	V/m
Broadcast	$2V/m$
ReduceScatter	V/m
Reduce	$2V/m$
AllReduce	$2V/m$

Cost when $MaxR$ and $MaxC$ are both at most V and $V >> m$

Communication Costs for mn-node 2D Supermesh S(m,n)

Communication	Cost
h-Reln, $h < mn/2$	$2V/m$
h-Reln, $h \geq mn/2$	nV/h
All-to-All	V/m
Copy	$2V * (1/(m+n) + 1/m(m+n))$
Scatter	$V * (1/(m+n) + 1/m(m+n))$
Gather	$V * (1/(m+n) + 1/m(m+n))$
Broadcast (1)	$V * (2m+n+1)/(m(m+n))$
Broadcast (2)	$V * (1 + 1/m) * (1/(m+n) + 1/n)$
ReduceScatter	V/m
Reduce	$V * (2m+n+1)/(m(m+n))$
AllReduce	$2V/m$

Communication Costs for Rack-Scale 144-node 2D Supermesh S(12,12)

Communication	Cost
h-Relation, $h < mn/2$	$0.167V$
h-Relation, $h \geq mn/2$	$12V/h$
h-Relation, $h = 108$	$0.111V$
All-to-All	$0.083V$
Copy	$0.087V$
Scatter	$0.045V$
Gather	$0.045V$
Broadcast (1)	$0.128V$
Broadcast (2)	$0.135V$
ReduceScatter	$0.083V$
Reduce	$0.128V$
AllReduce	$0.167V$

Size of V could be less than a MB or hundreds of MB.

All general and collective communications can be done 5x-20x faster than the time to send a vector of size V across a single link!

On many HPC architectures with this number of nodes, cost might be 8V or more due to hops, congestion etc.

Supermeshes can potentially be 10x-100x faster than current HPC interconnects, with the same link speeds.

Thank You!

HUAWEI

ZURICH RESEARCH CENTER