Itasca's IB Fabric

1074 nodes

8536 cores

25.1 TB memory

140 TB storage

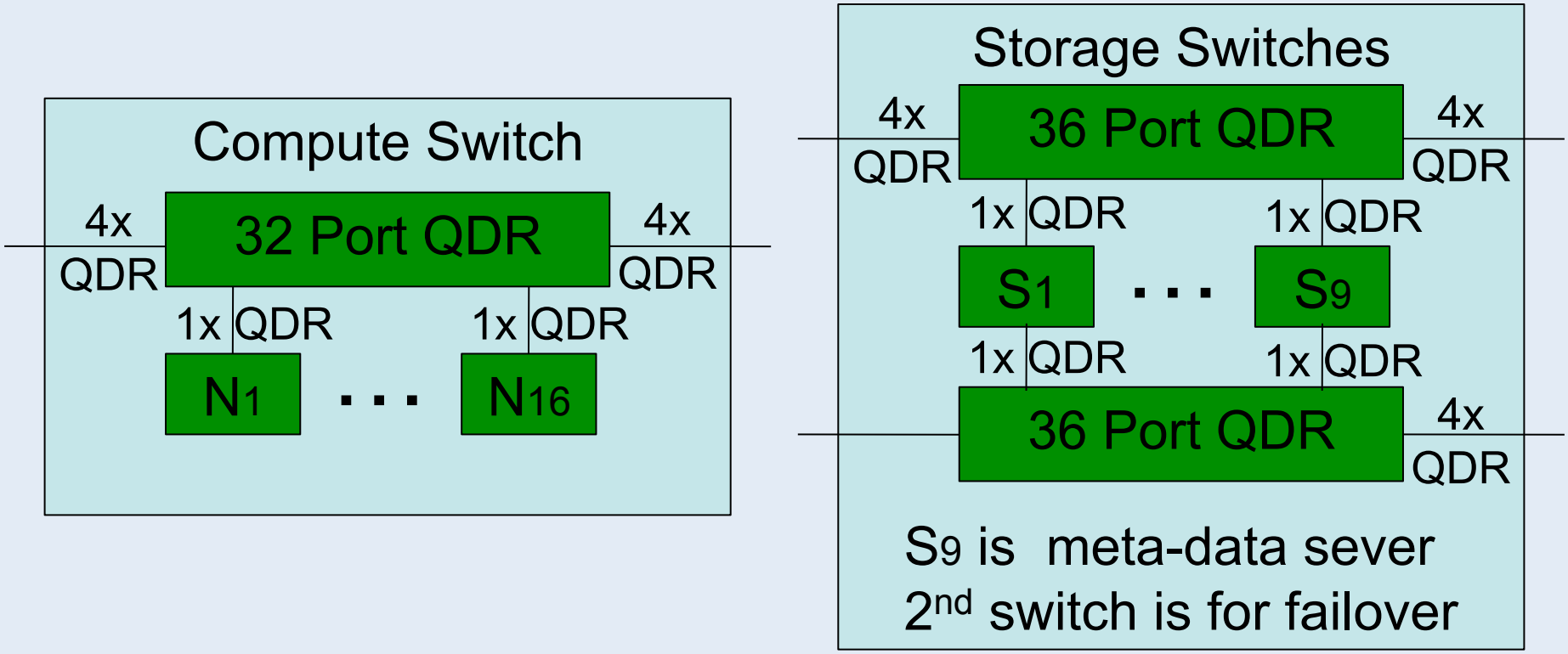
How is it all connected?

Scalability and performance?

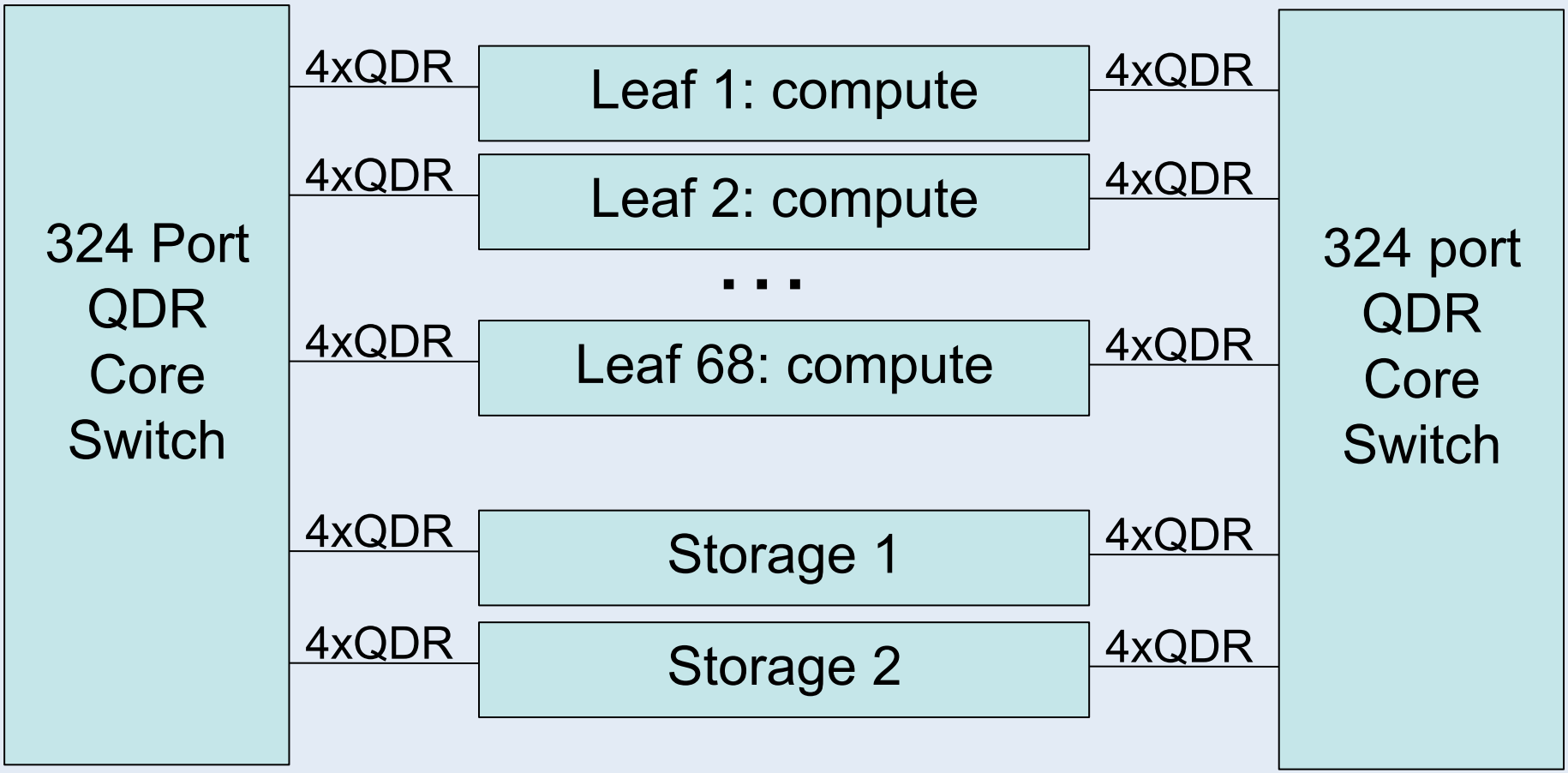
Itasca's Deep IB Hierarchy

- 8 cores per node
  - 2x Nehalem processors, 4 cores per proc.
  - 24 GB GB shared memory
- 16 nodes per "leaf switch"
  - QDR IB to each node
- 68 leaf switches connect to 2 directors
  - 4x QDR between each leaf and director switch

Compute & Storage Switches



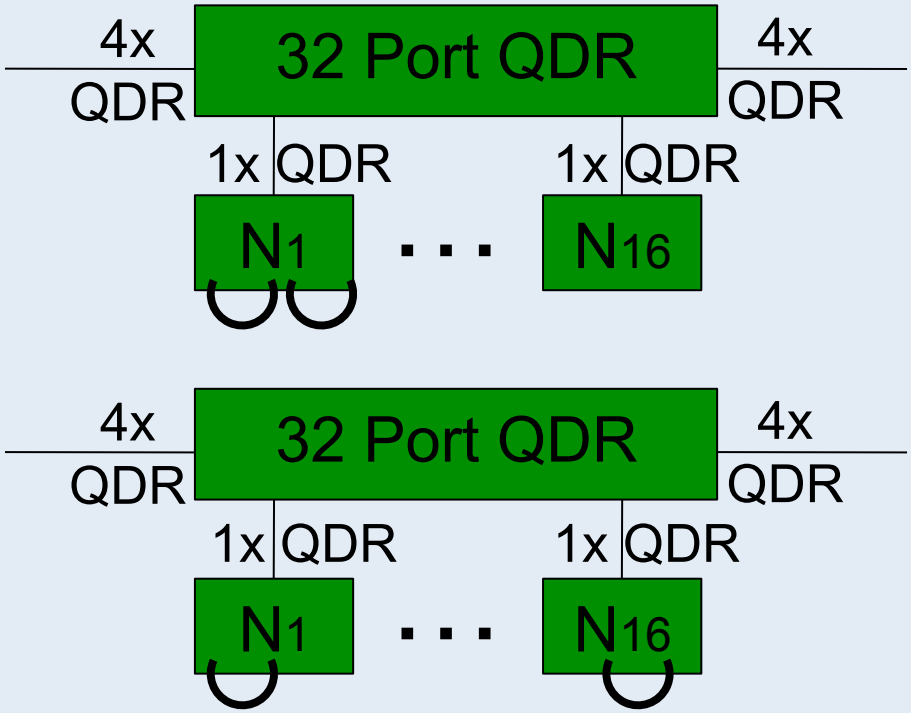
Leaf & Director Switches



IB Performance Test

- Ping-pong between random pairs of MPI ranks
  - Measure time to do many (100) ping-pongs
  - Use large messages to measure bandwidth
  - Use blocking mpi_send & mpi_recv
- Bandwidth = total data per node / total time
- Collect & report timing data on rank 0

Within a Node
No IB Contention



Message size: 1MB

# of Iterations: 100

MPI ranks: 256

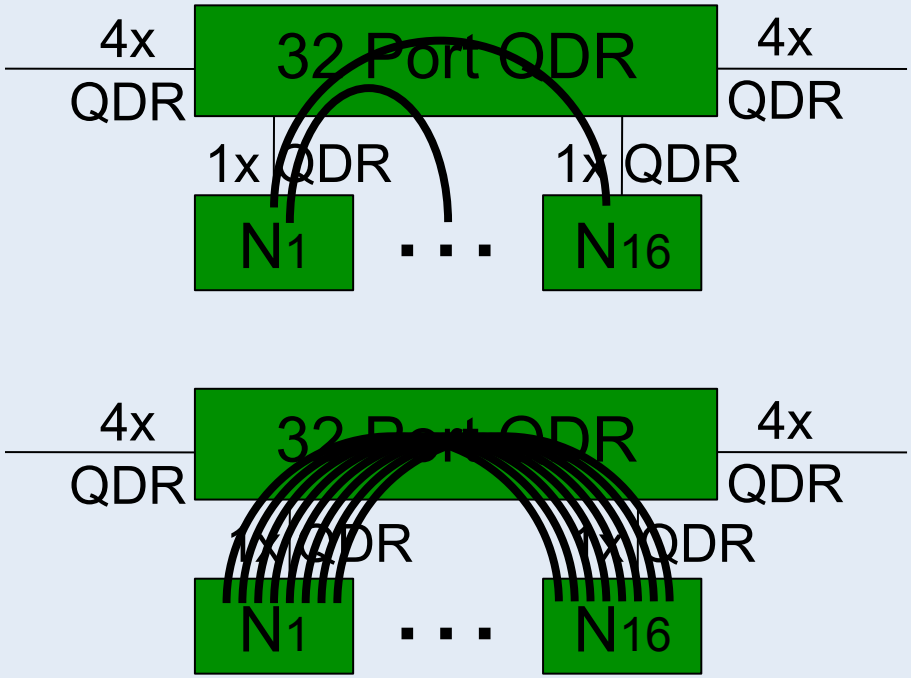
MPI ranks per node: 8

Max	Average	Min	[MB/sec]
3773.94	482.49	390.49	Overall

3773.94	3426.51	2644.41	Same Node
783.64	489.96	424.83	Same Leaf
934.82	454.64	390.49	Diff Leafs

Does not use IB

Between Nodes 8:1 contention



Message size: 1MB

# of Iterations: 100

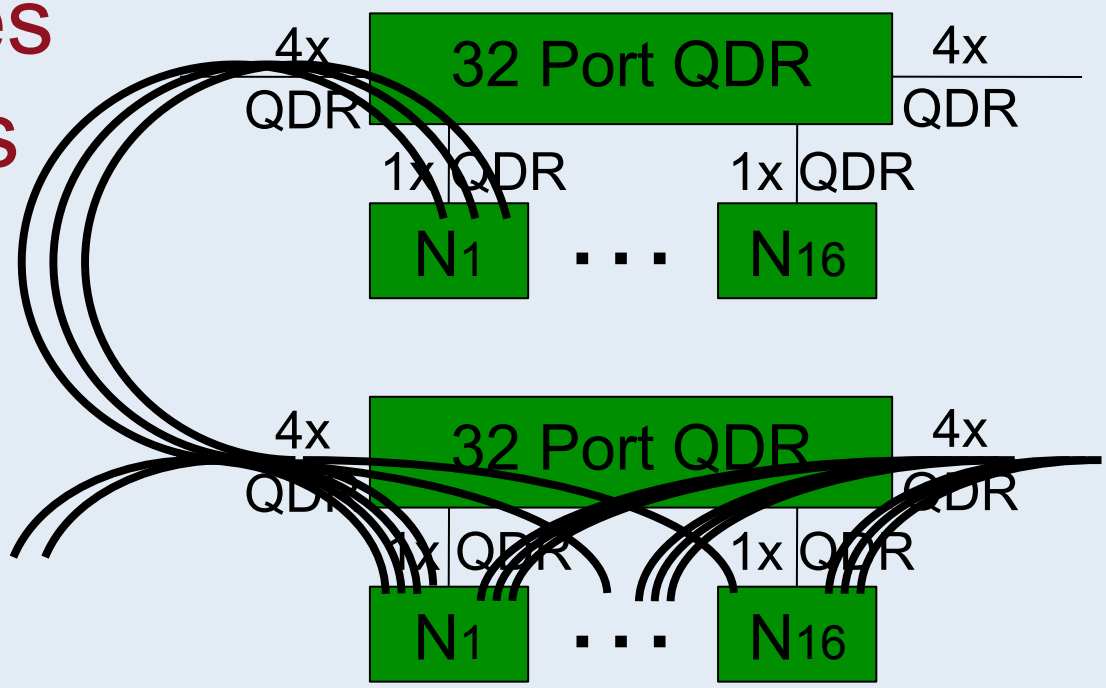
MPI ranks: 256

MPI ranks per node: 8

Max	Average	Min	[MB/sec]
3773.94	482.49	390.49	Overall
3773.94	3426.51	2644.41	Same Node
783.64	489.96	424.83	Same Leaf
934.82	454.64	390.49	Diff Leafs

As much as 8 fold contention

Between Switches 8 ports:16 nodes



Message size: 1MB

# of Iterations: 100

MPI ranks: 256

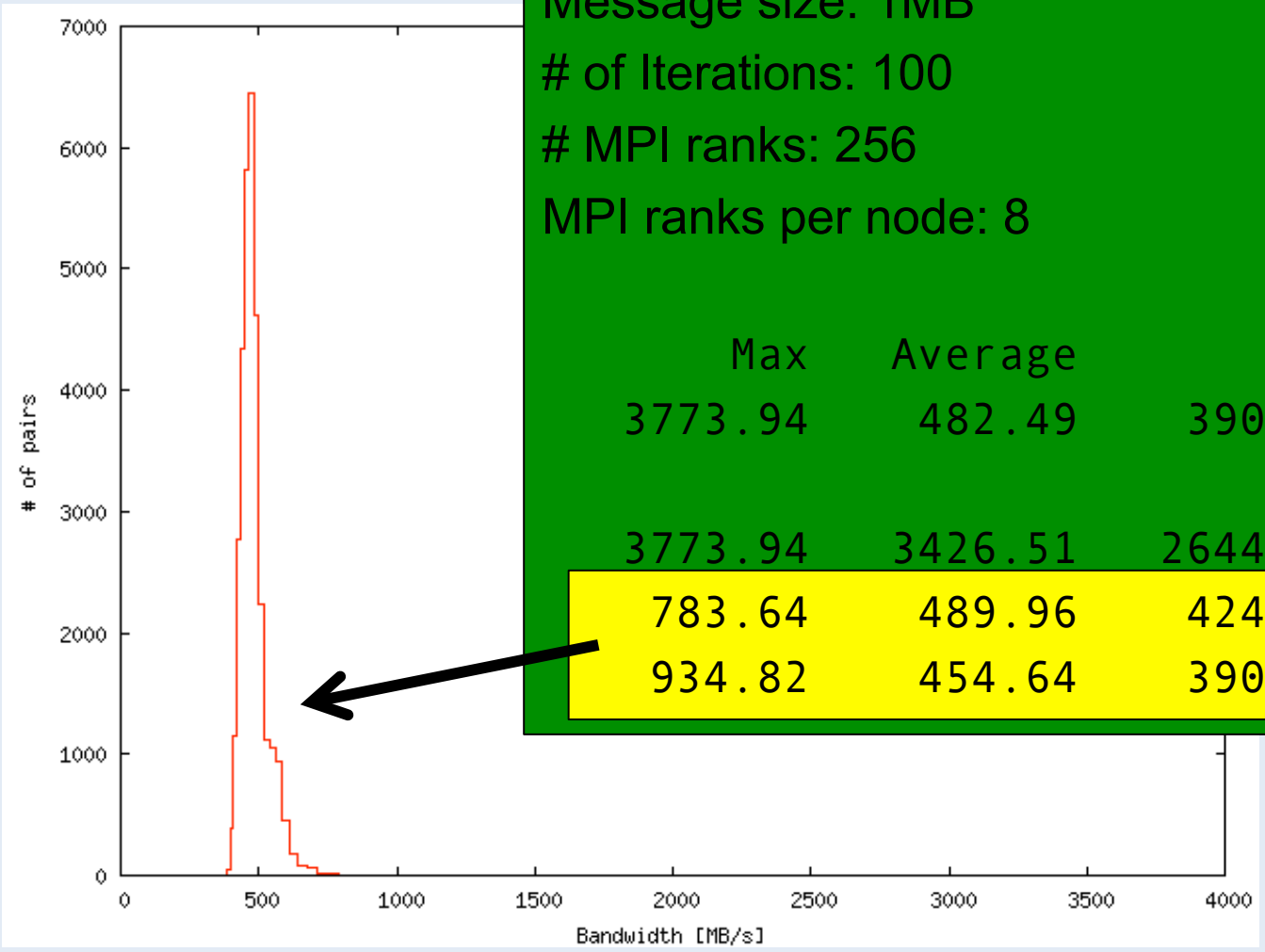
MPI ranks per node: 8

Max	Average	Min	[MB/sec]
3773.94	482.49	390.49	Overall
3773.94	3426.51	2644.41	Same Node
783.64	489.96	424.83	Same Leaf
934.82	454.64	390.49	Diff Leafs

Extreme variability due to random set of pairs.

Numbers are typical

Random Pairs, 8 Ranks per Node



Message size: 1MB

# of Iterations: 100

MPI ranks: 256

MPI ranks per node: 8

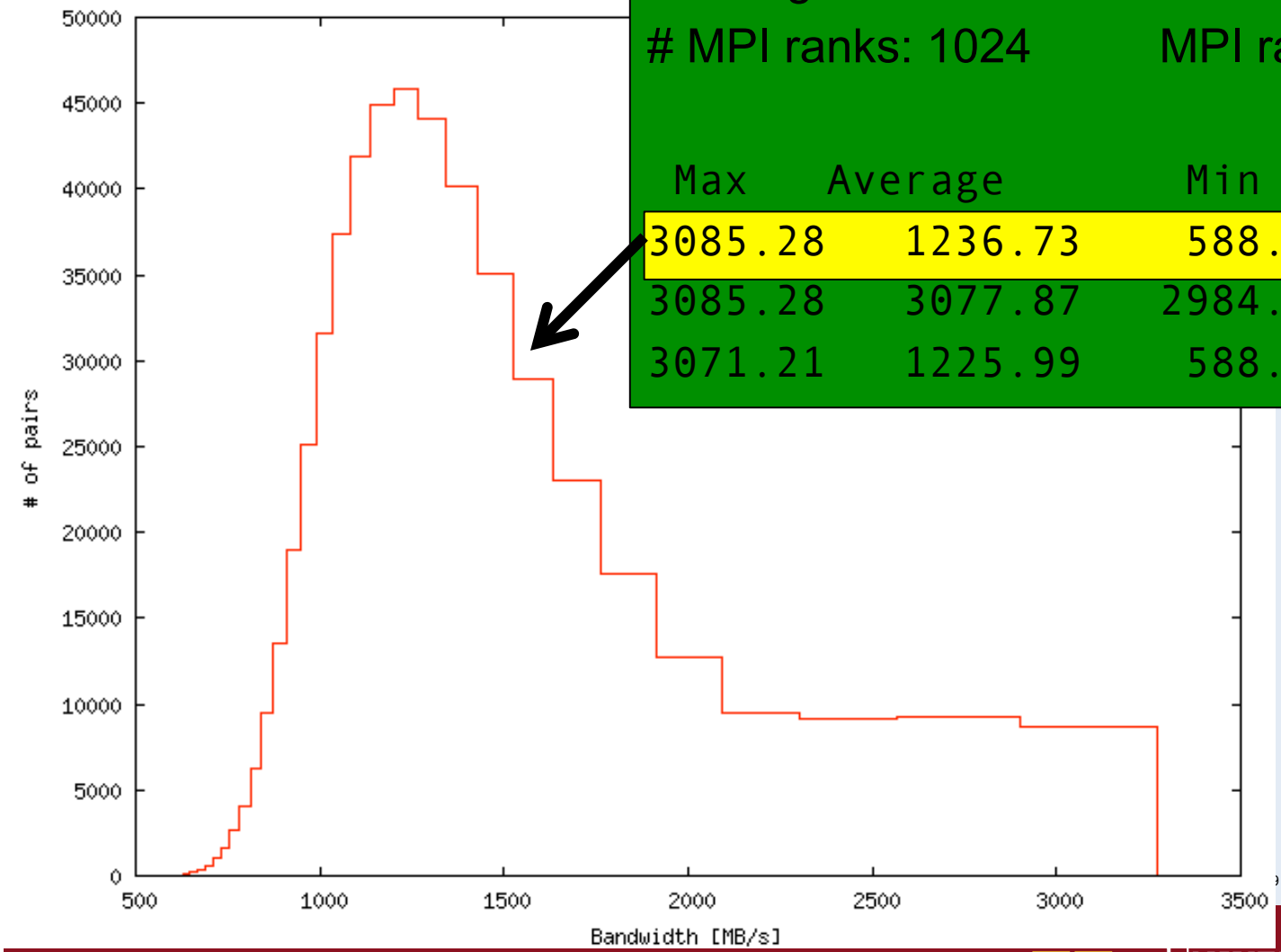
Max	Average	Min	[MB/sec]
3773.94	482.49	390.49	Overall
3773.94	3426.51	2644.41	Same Node
783.64	489.96	424.83	Same Leaf
934.82	454.64	390.49	Diff Leafs

Random Pairs: 1 Rank per Node

Message size: 1MB # of Iterations: 100

MPI ranks: 1024 MPI ranks per node: 1

Max	Average	Min	[MB/sec]
3085.28	1236.73	588.25	Overall
3085.28	3077.87	2984.23	Same Leaf
3071.21	1225.99	588.25	Diff Leafs



Same Leaf: Fairly non-blocking

Director Switch Contention

QDR: 4 GB/s @ 75% → 3071 MB/s

Same leaf with 1 rank per node

Max	Average	Min	
3085.28	3077.87	2984.23 [MB/s]	***OK***

Different leafs with up to 16 nodes using 4+4 ports

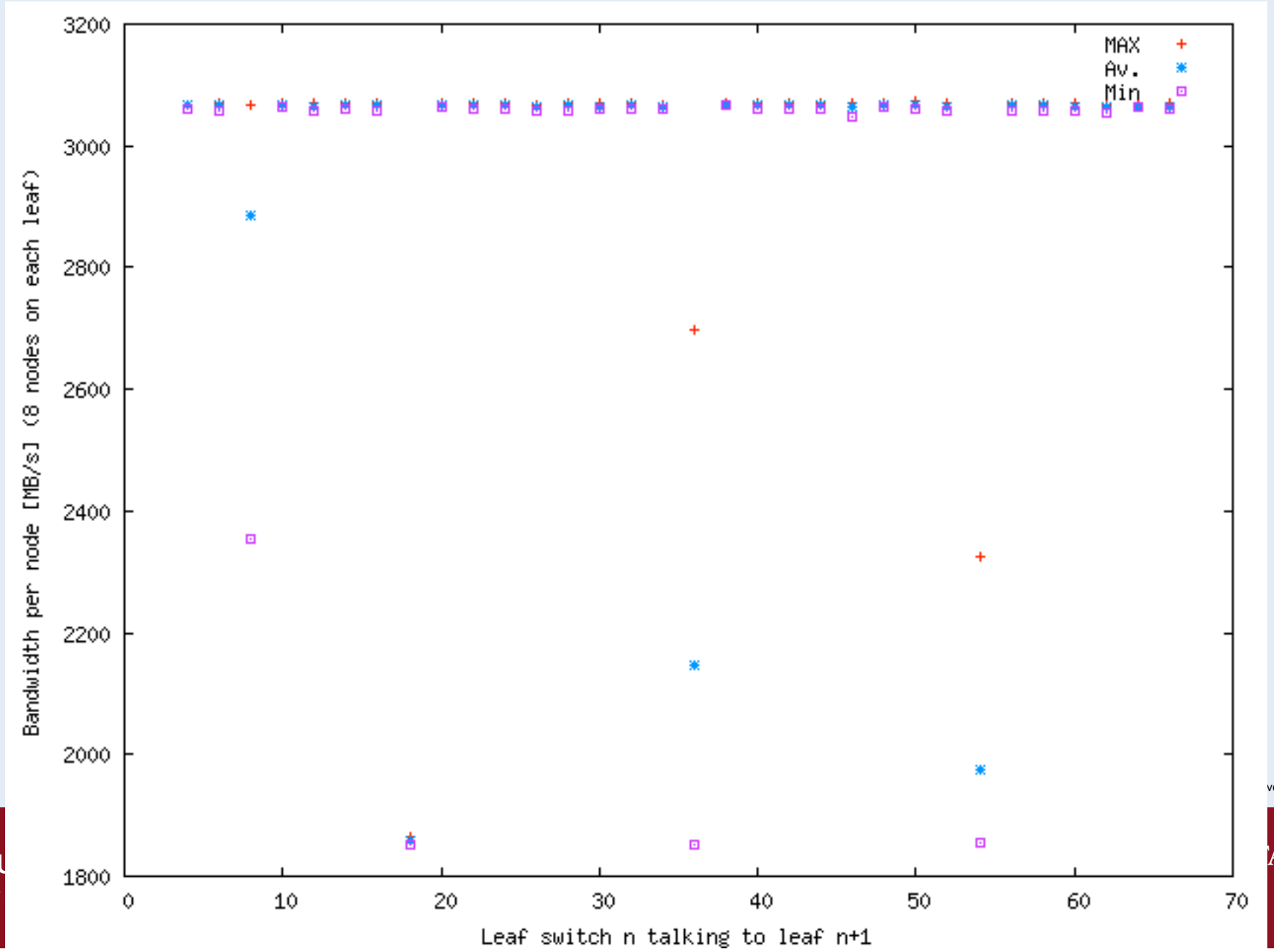
Max	Average	Min	
3071.21	1225.99	588.25 [MB/s]	***Expected ~1500 MB/s***

Expected no more that 2:1 contention from leaf to director switch

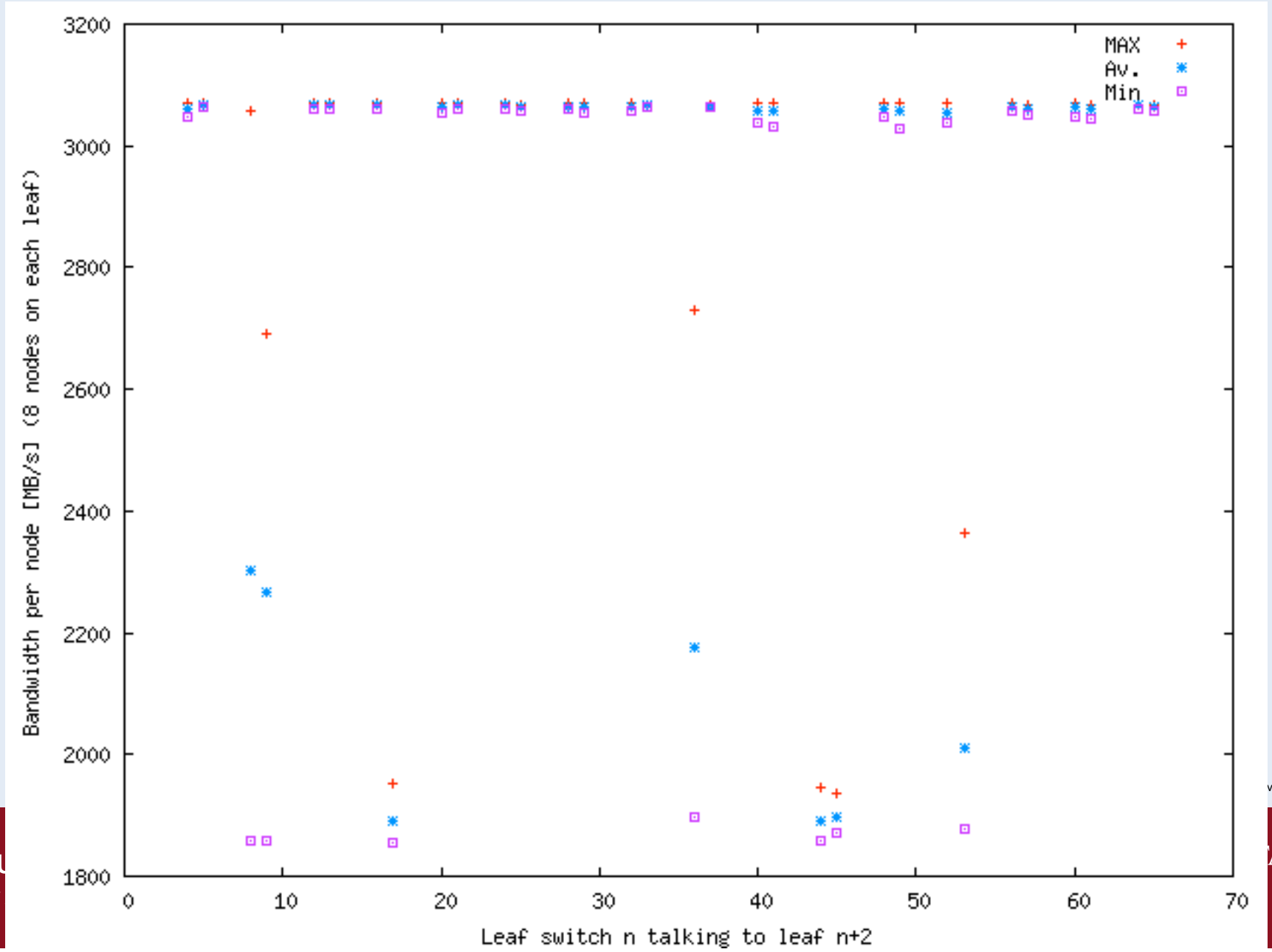
Leaf to Leaf Bandwidth Test

- 1 MPI rank per node
- 8 nodes per leaf: on for each director port
  - Static routing: 1st 8 nodes evenly distributed
- Select offset between leafs
  - N-th node on one leaf only talks to N-th node on any other leaf.
  - Offset of 1: leafs 1-2, 3-4, 5-6, 7-8 …
  - Offset of 2: leafs 1-3, 2-4, 5-7, 6-8, …
  - Offset of 4: leafs 1-5, 2-6,3-7,4-8, 9-13….

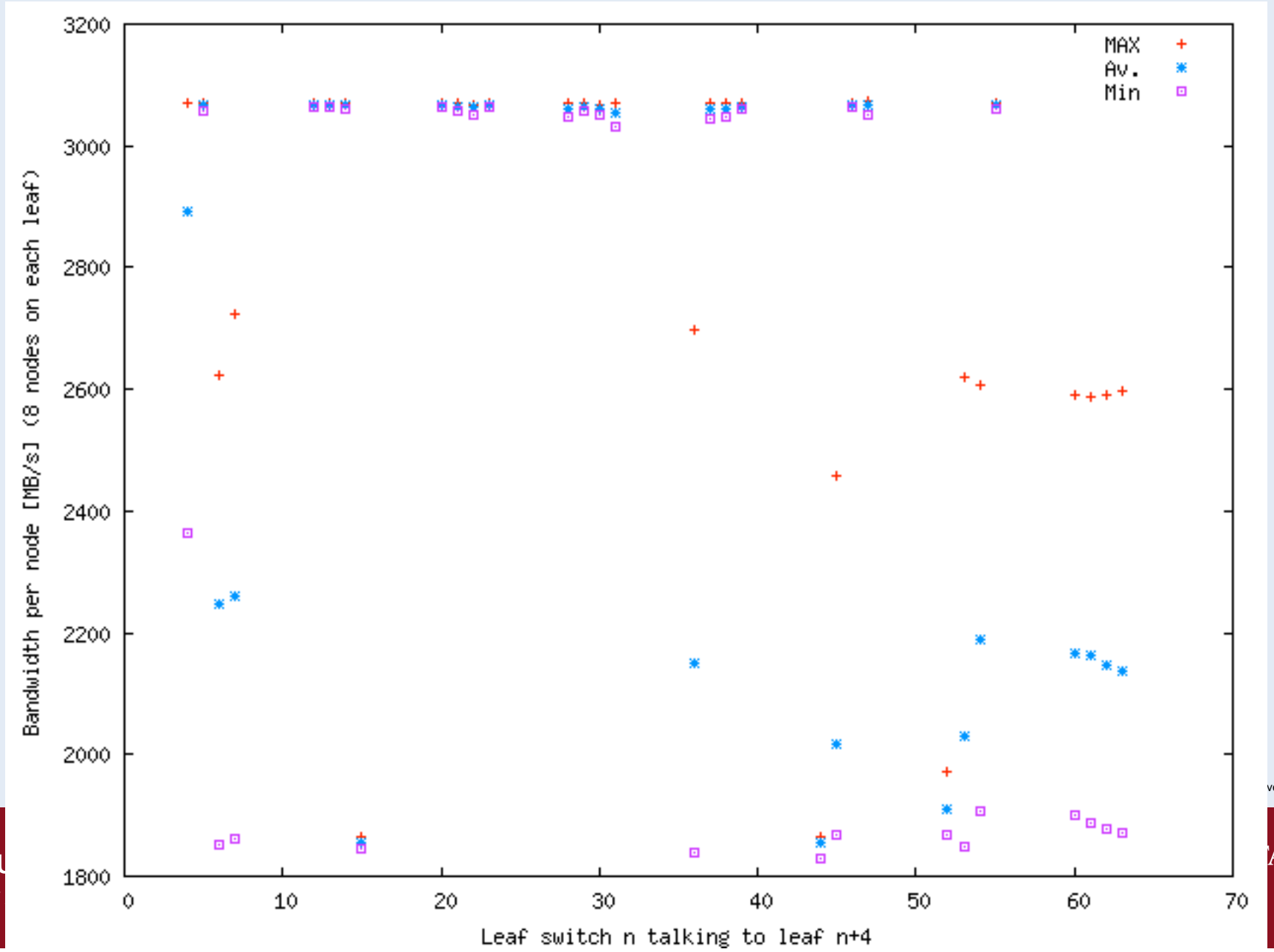
Leaf-to-leaf: offset=1



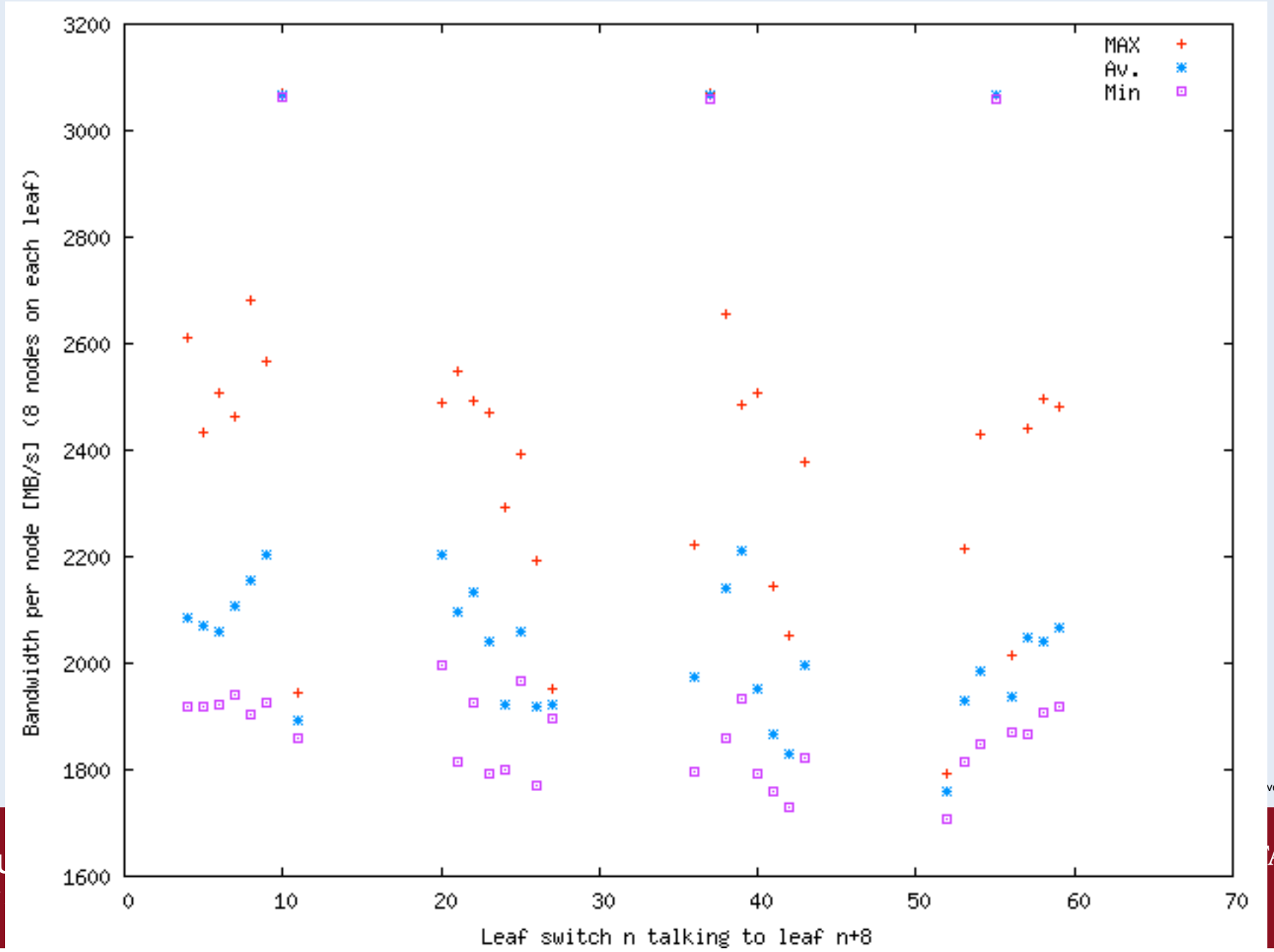
Leaf-to-leaf: offset=2



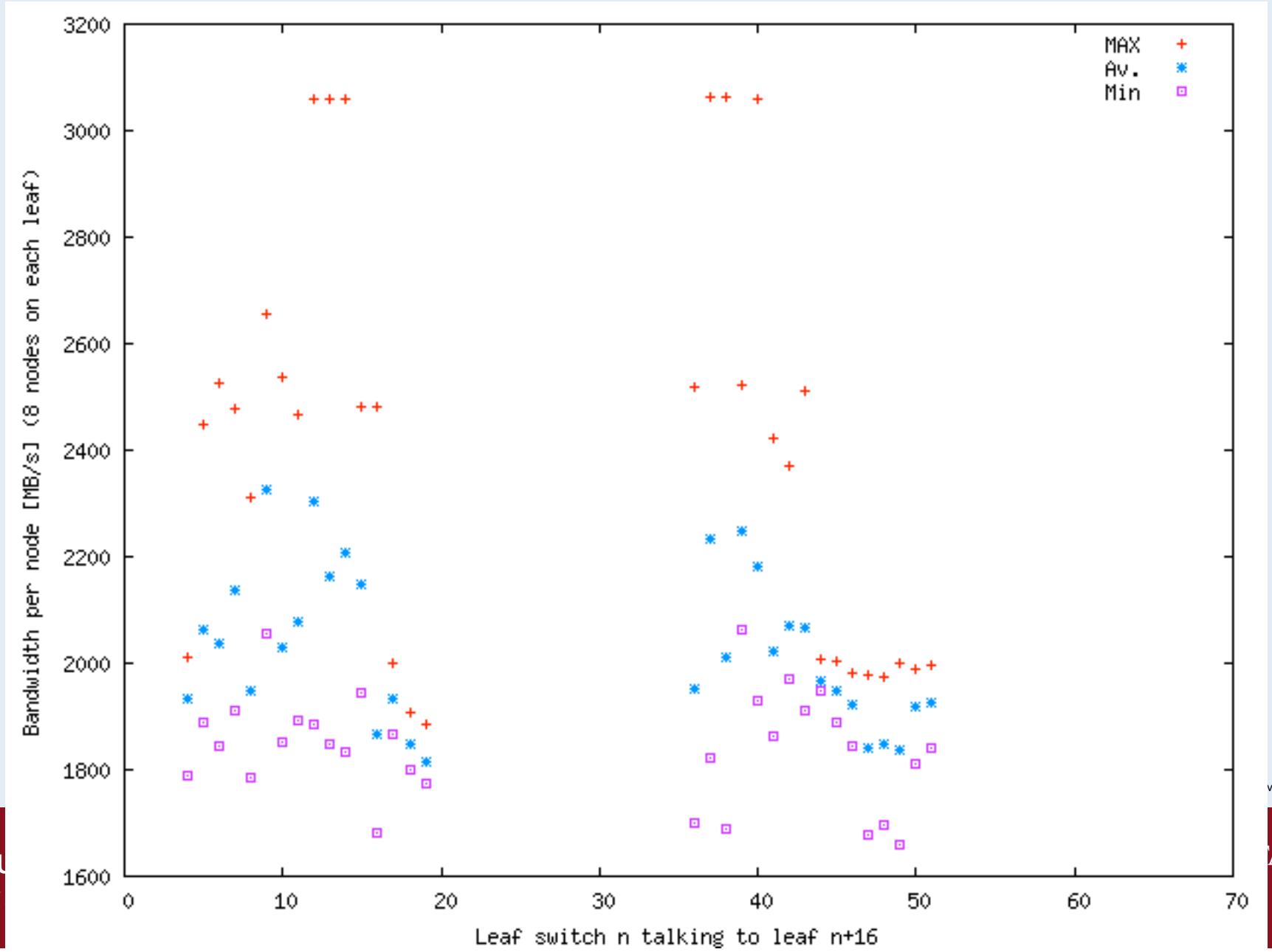
Leaf-to-leaf: offset=4



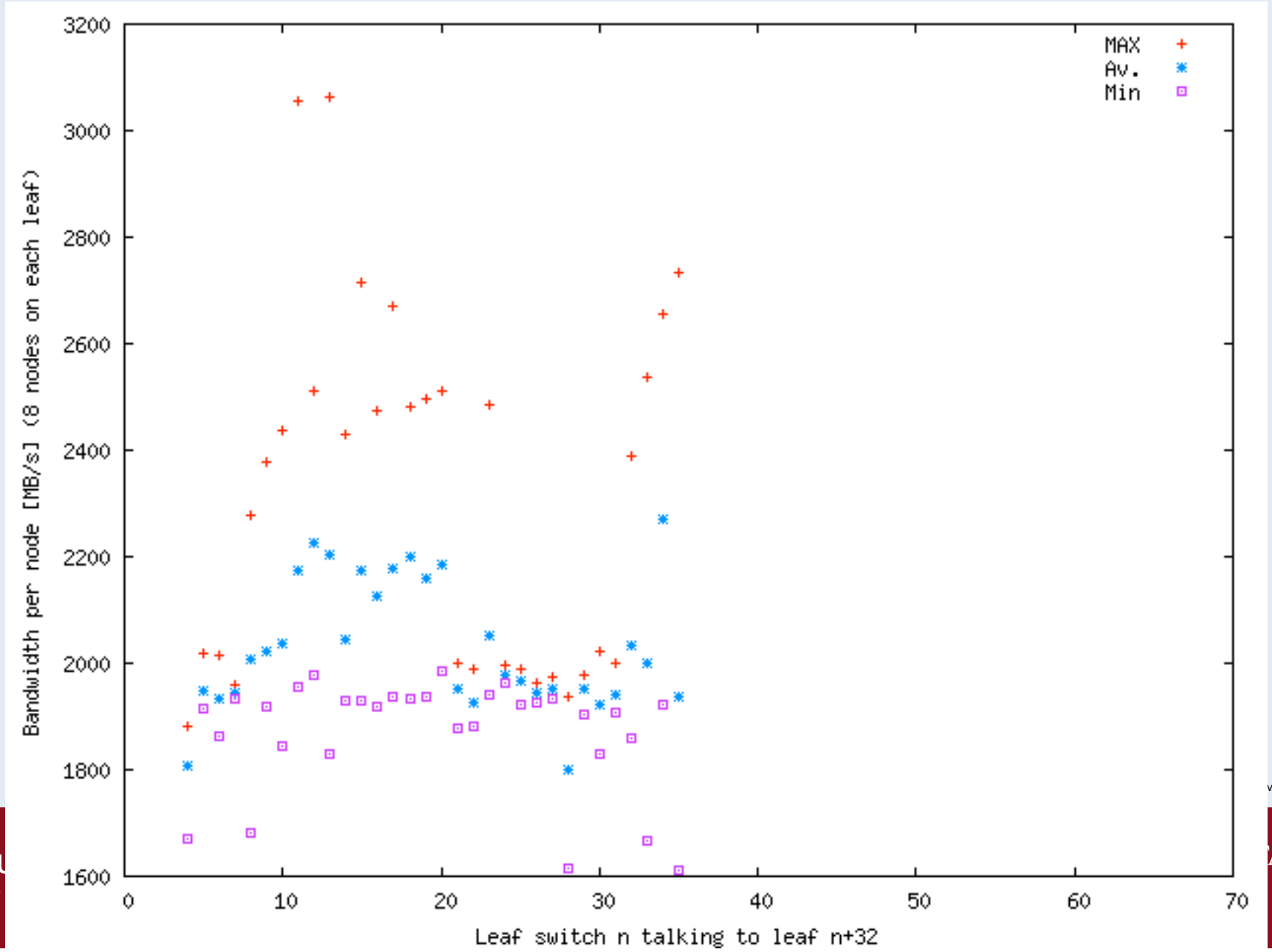
Leaf-to-leaf: offset=8



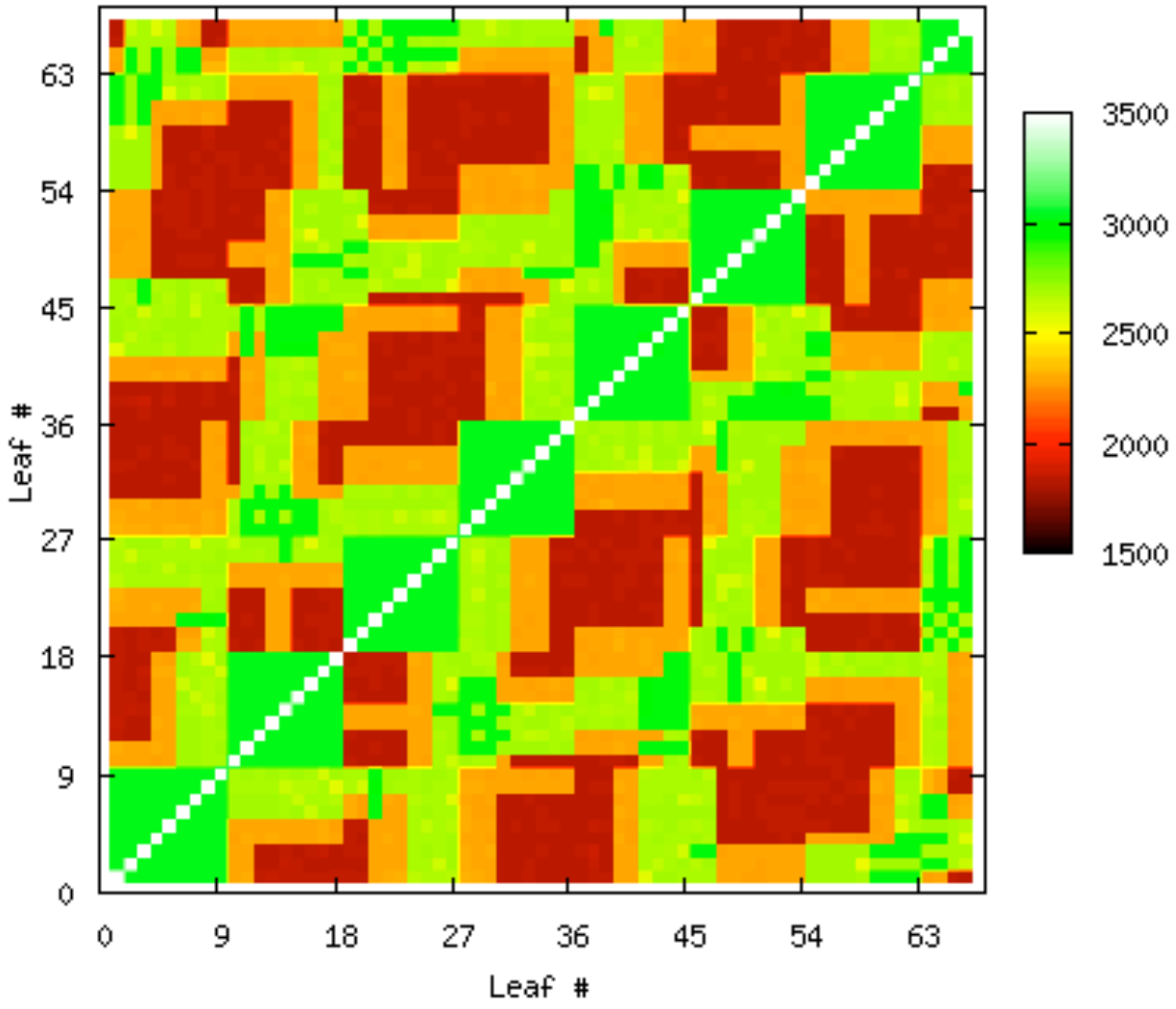
Leaf-to-leaf: offset=16



Leaf-to-leaf: offset=32



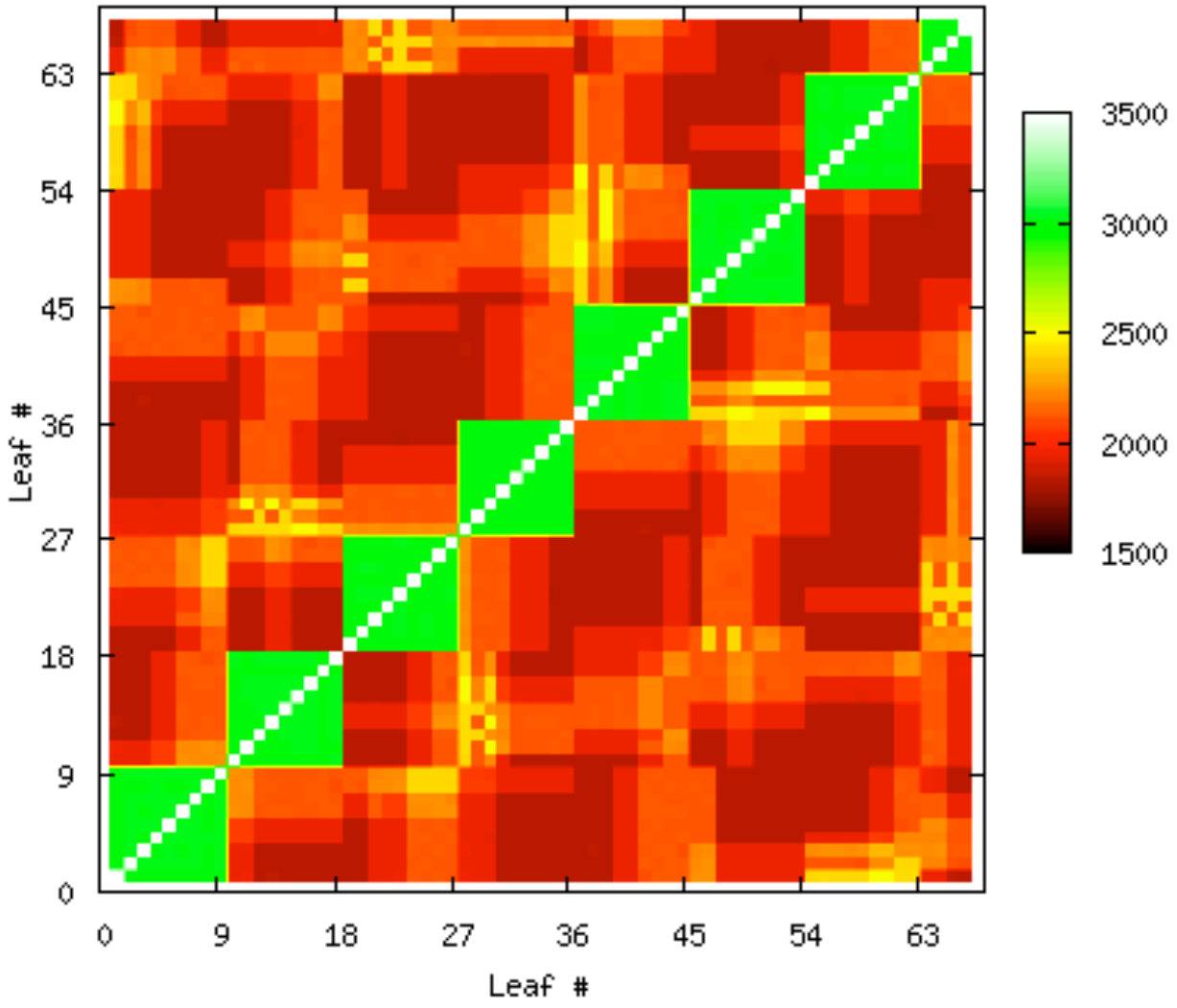
All Pairs: Max Bandwidths



- Range of bandwidths:
  Max: 3035 MB/s
  Min: 1826 MB/s
- Shown here: Max bandwidth out of the 8 nodes on one leaf talking to 8 node on a different leaf

Surprising amount of variation

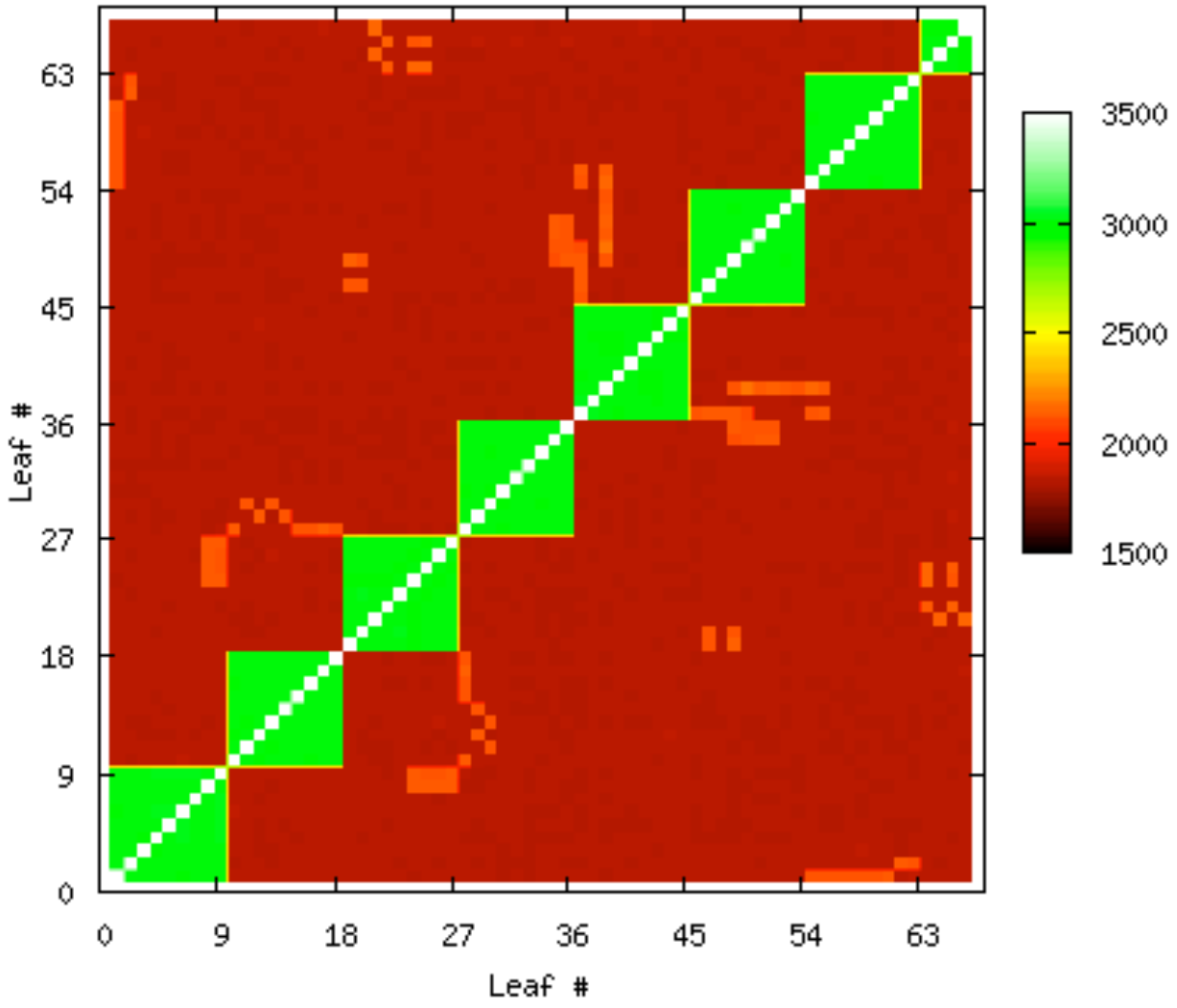
All Pairs: Average Bandwidths



- Range of bandwidths:
  Max: 3035 MB/s
  Min: 1826 MB/s
- Shown here: Average bandwidth out of the 8 nodes on one leaf talking to 8 node on a different leaf
- Contiguous set of leafs show good communication.

A pattern emerges

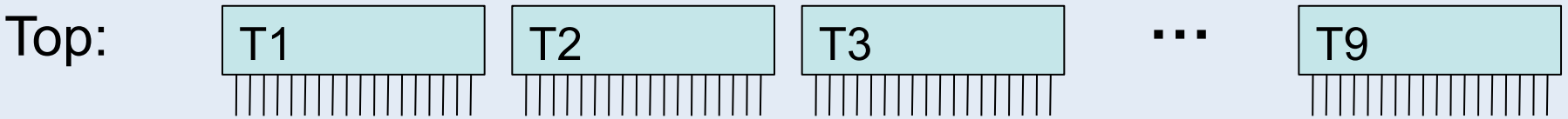
All Pairs: Min Bandwidths



- Range of bandwidths:
  Max: 3035 MB/s
  Min: 1826 MB/s
- 8 nodes on a leaf
- 8 QDR ports to director switches
- Should have been NO contention.
  Expected: ~3000 MB/s

Need to understand director switches

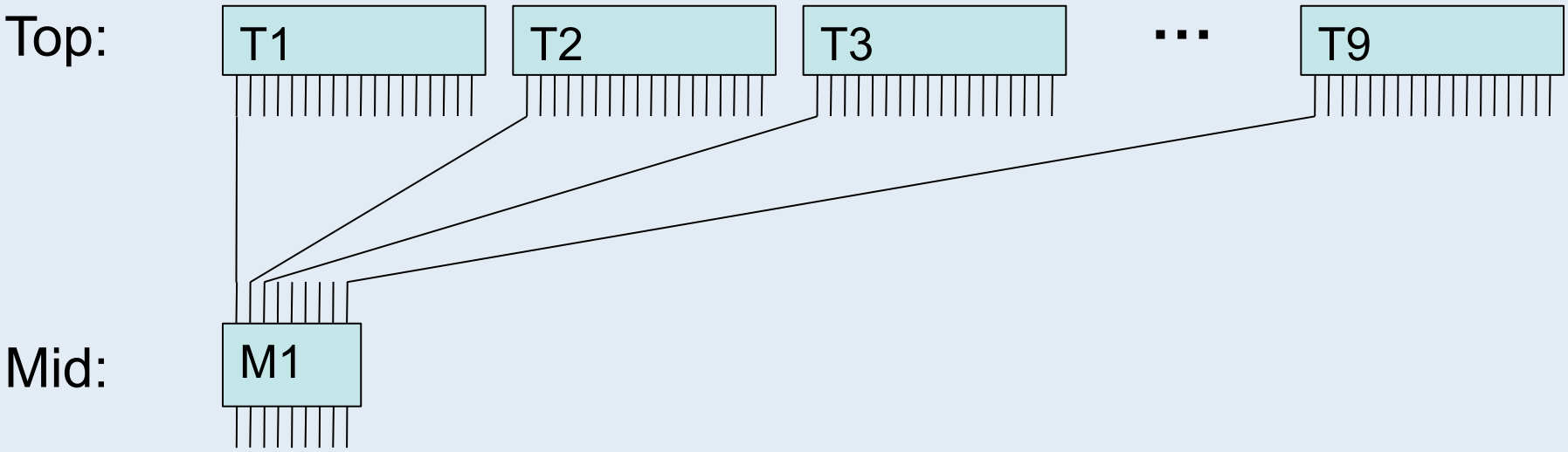
Director Switch: Top Level



- Top level switches: 36 QDR ports each
- All 36 ports point “down to mid level switches
- Total of 9 top level switches
- 18 of the 36 ports point to top level switches

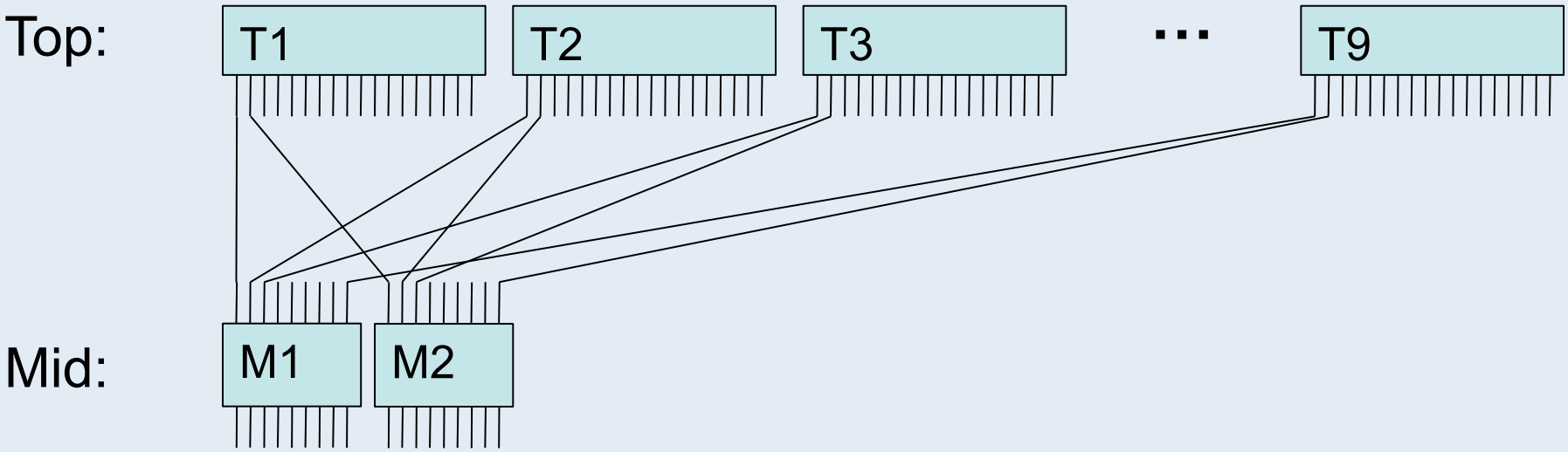
(Each line here represents a pair of 2 QDR lines)

Director Switch: Mid Level



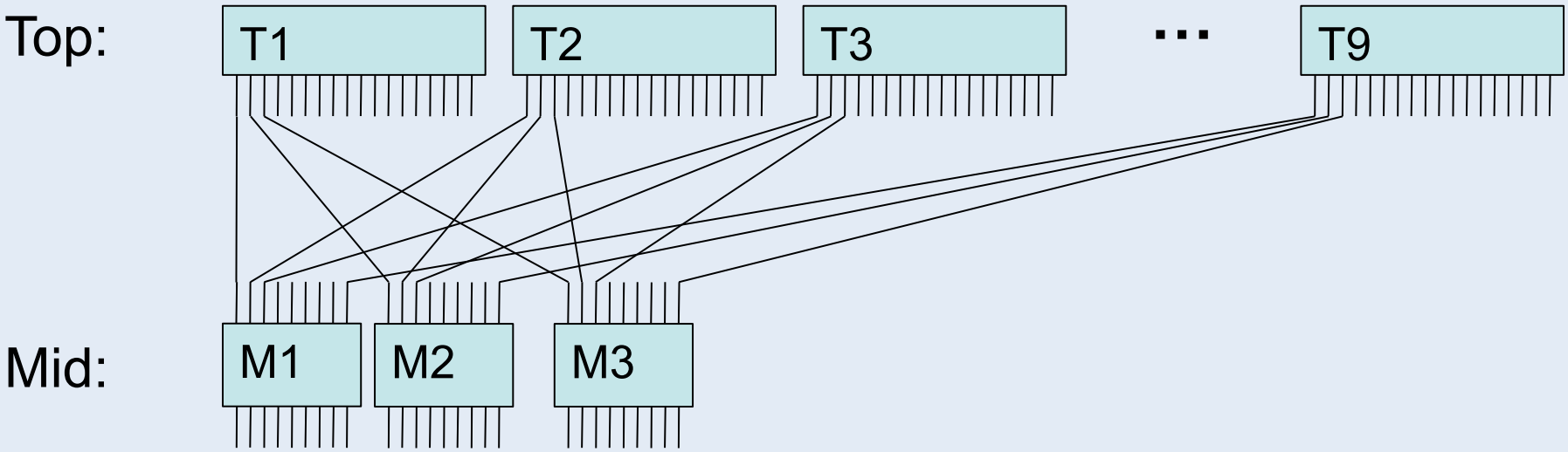
- Mid level switches: 36 QDR ports each (same as top level)
- Each line here represents 2 QDR lines
- 18 of the 36 ports point to top level switches

Director Switch: Mid Level



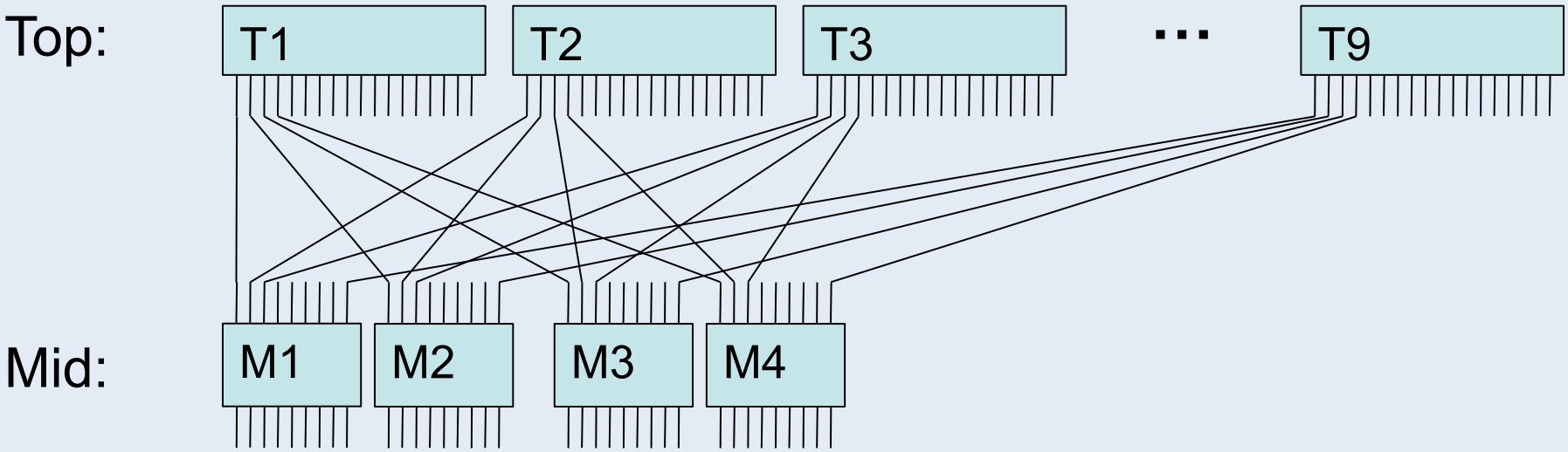
- Mid level switches: 36 QDR ports each (same as top level)
- Each line here represents 2 QDR lines
- 18 of the 36 ports point to top level switches
- Every mid level switch has 2 QDR lines to every top level

Director Switch: Mid Level



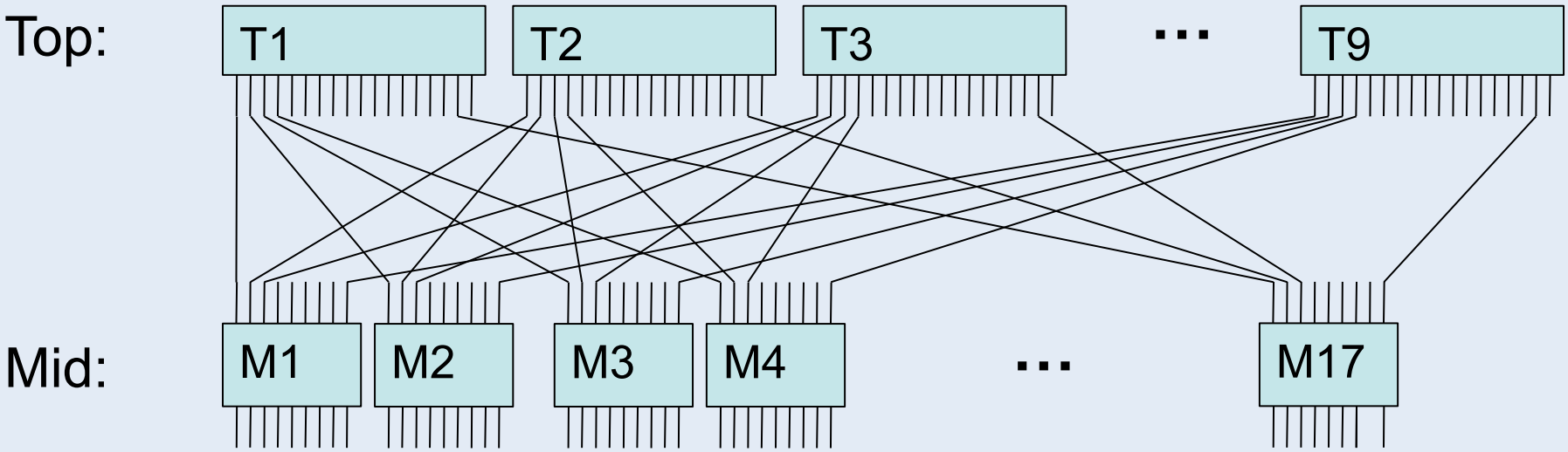
- Mid level switches: 36 QDR ports each (same as top level)
- Each line here represents 2 QDR lines
- 18 of the 36 ports point to top level switches
- Every mid level switch has 2 QDR lines to every top level

Director Switch: Mid Level



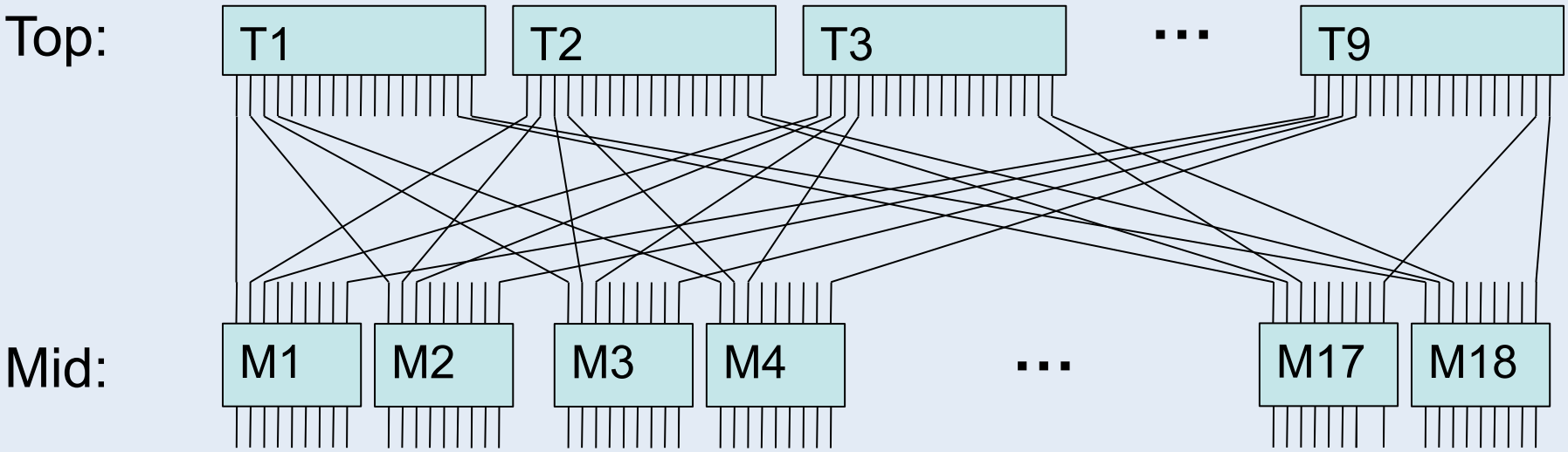
- Mid level switches: 36 QDR ports each (same as top level)
- Each line here represents 2 QDR lines
- 18 of the 36 ports point to top level switches
- Every mid level switch has 2 QDR lines to every top level

Director Switch: Mid Level



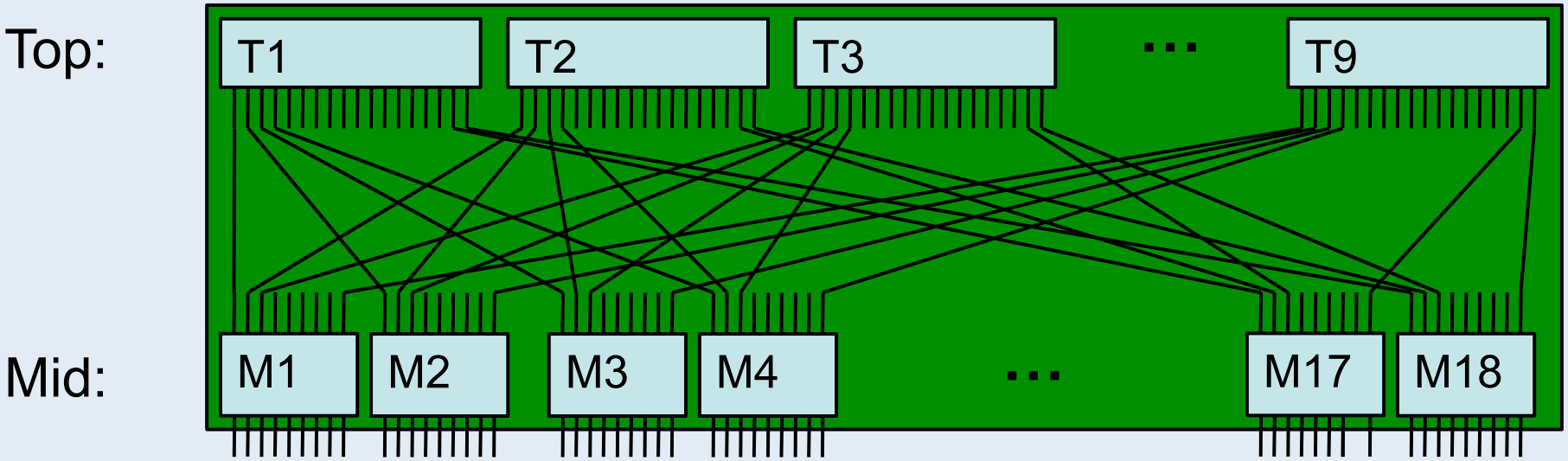
- Mid level switches: 36 QDR ports each (same as top level)
- Each line here represents 2 QDR lines
- 18 of the 36 ports point to top level switches
- Every mid level switch has 2 QDR lines to every top level

Director Switch: Mid Level



- Mid level switches: 36 QDR ports each (same as top level)
- Each line here represents 2 QDR lines
- 18 of the 36 ports point to top level switches
- Every mid level switch has 2 QDR lines to every top level
- A total of 18 mid level switches: 18*18 = 324 = 9*36

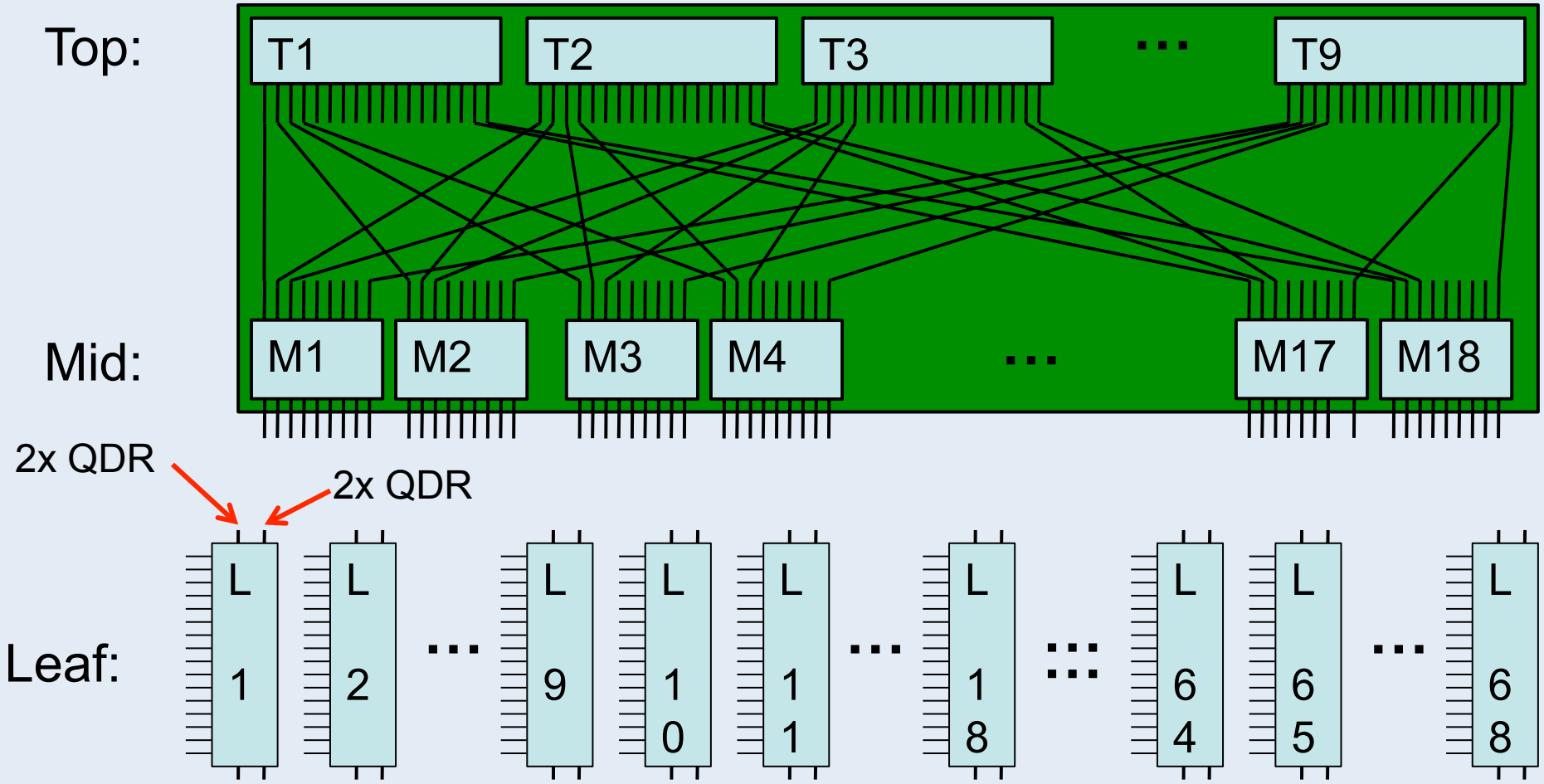
Itasca Director Switch



324 QDR ports
18 ports from each mid level switch
Each line here represents 2 ports or 2 QDR lines

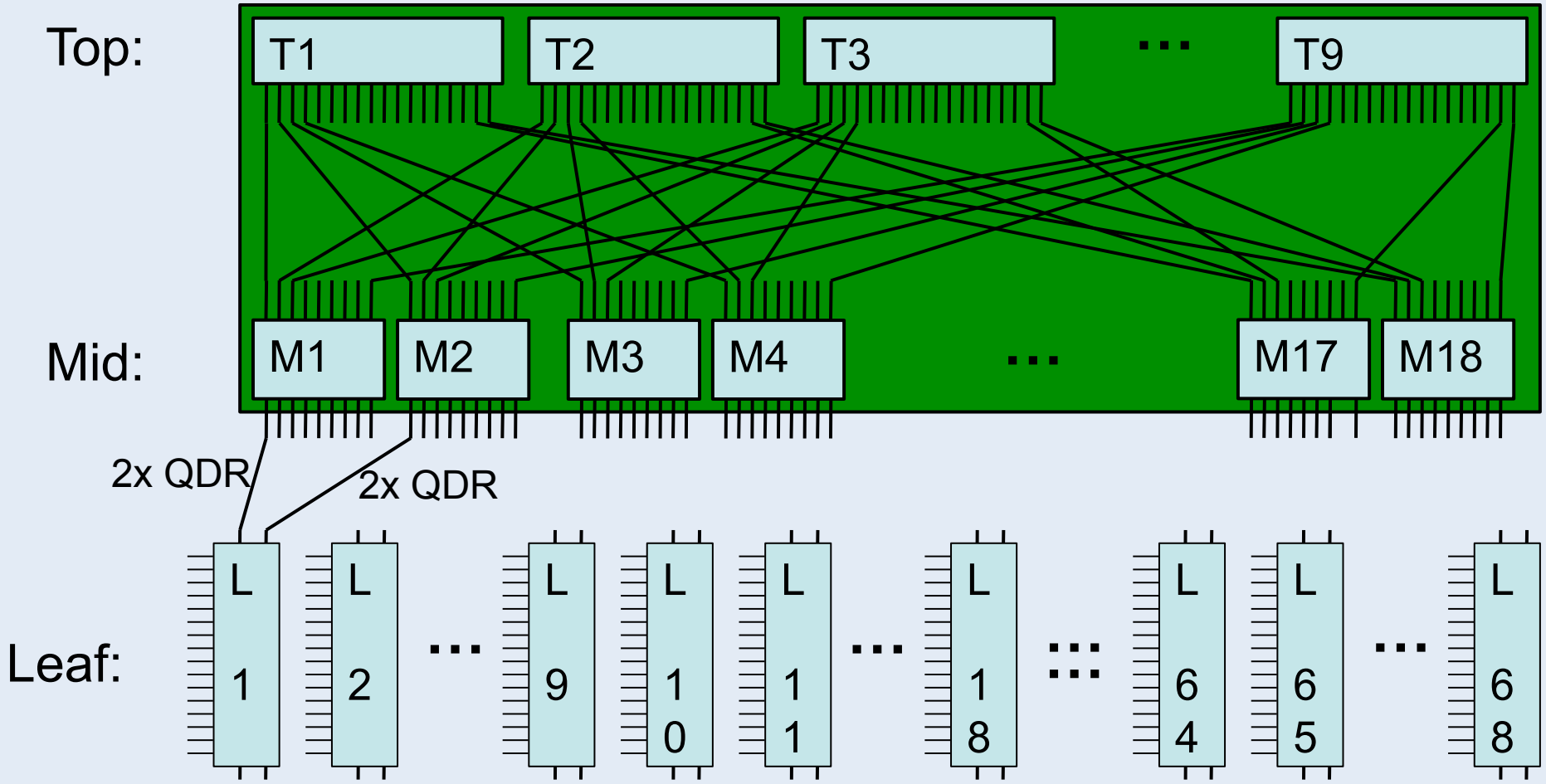
Itasca has two of these director switches
Each director switch is a “Fat Tree”

Itasca Leaf Switches



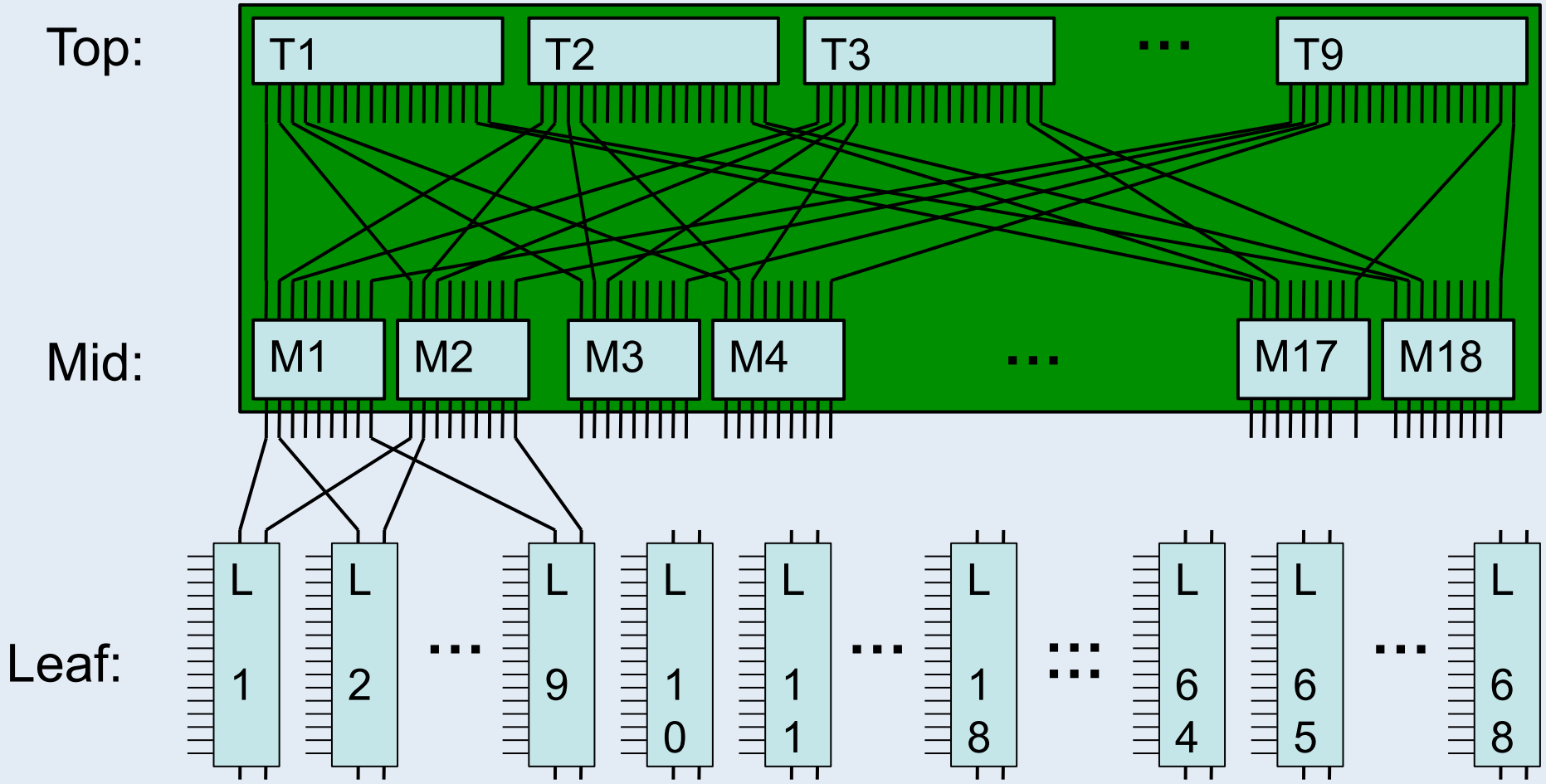
67 leaf switches – 16 compute nodes per leaf switch

Leaf to Director Connection



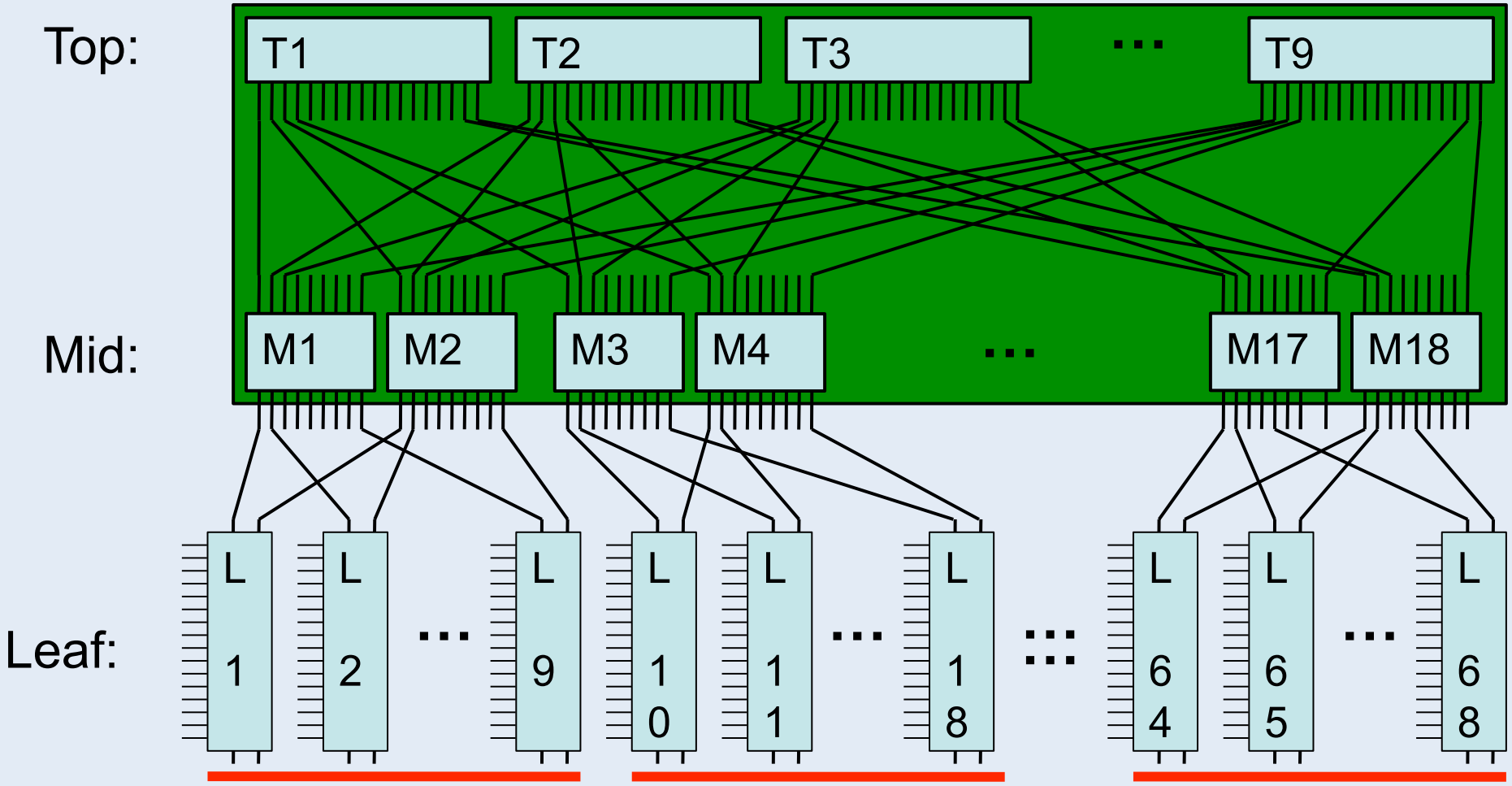
Each leaf switch has 2 QDR lines to each of 2 mid level switches

Leaf to Director Connection



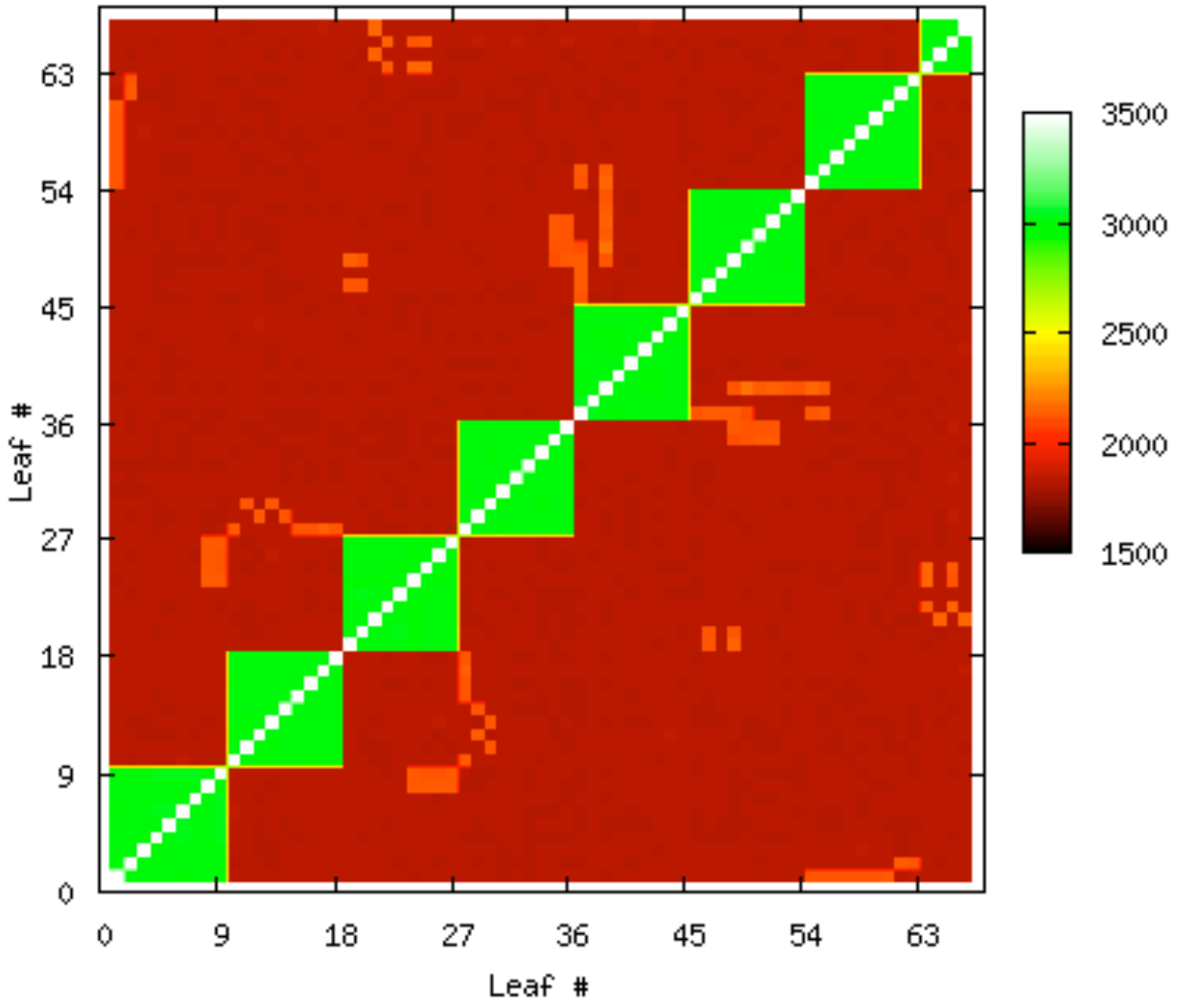
Sets of 9 leafs connected to same 2 mid level switches

Leaf to Director Connection



"Near" sets of 9 leafs connected to same 2 mid level switches

All Pairs: Min Bandwidths



- Range of bandwidths:
  Max: 3035 MB/s
  Min: 1826 MB/s
- Blocks of 3000 MB/s align with sets of 9 “near” leafs.
- All “far” leafs (not in set of 9 “near”) can have an extra factor of 2 contention

Fat Tree is not the same as Full Crossbar

Comparison with Blade and Calhoun

- Random pairs
- 1 rank per node
- Message size=1 MB
- Bandwidths [MB/s]

	Max	Average	Min	# Nodes in Test
Blade	920	767	661	4
Calhoun	1005	844	611	64
Itasca	3085	1236	588	1024

STATIC ROUTINING in Itasca's IB switches
In principle 16 nodes could all go through one QDR link.
More testing/tuning of director switches is needed.

Applications Can Scale Well On Itasca

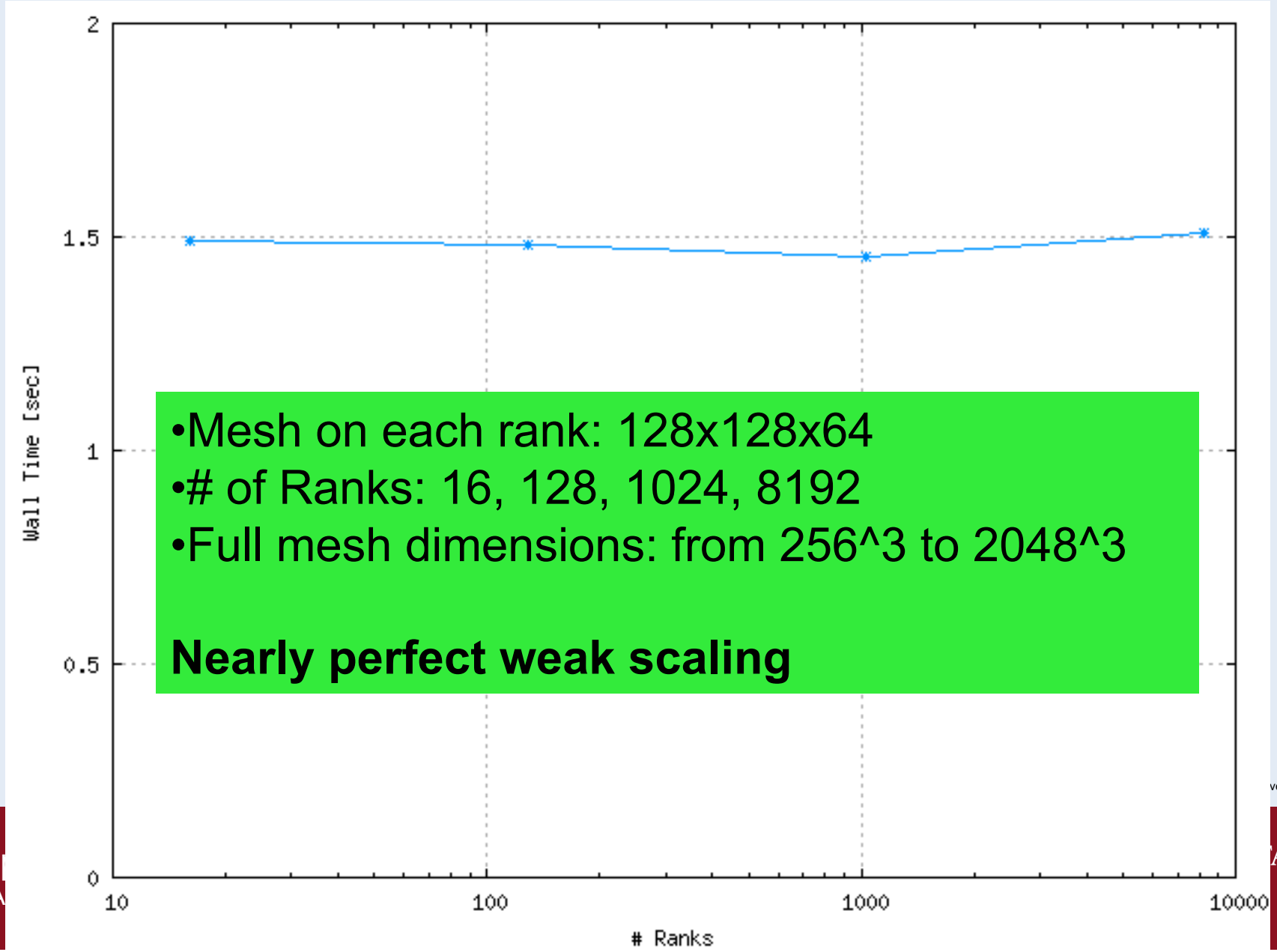
Test code: Isothermal MHD using TVD scheme

- Uniform mesh
- 3D domain decomposition
- Nearest neighbor communication
- Communication overlapped with computation

Scaling tests from 16 to 8192 compute ranks

- Weak scaling: problem size grows with number of ranks
- Strong scaling: problem size fixed

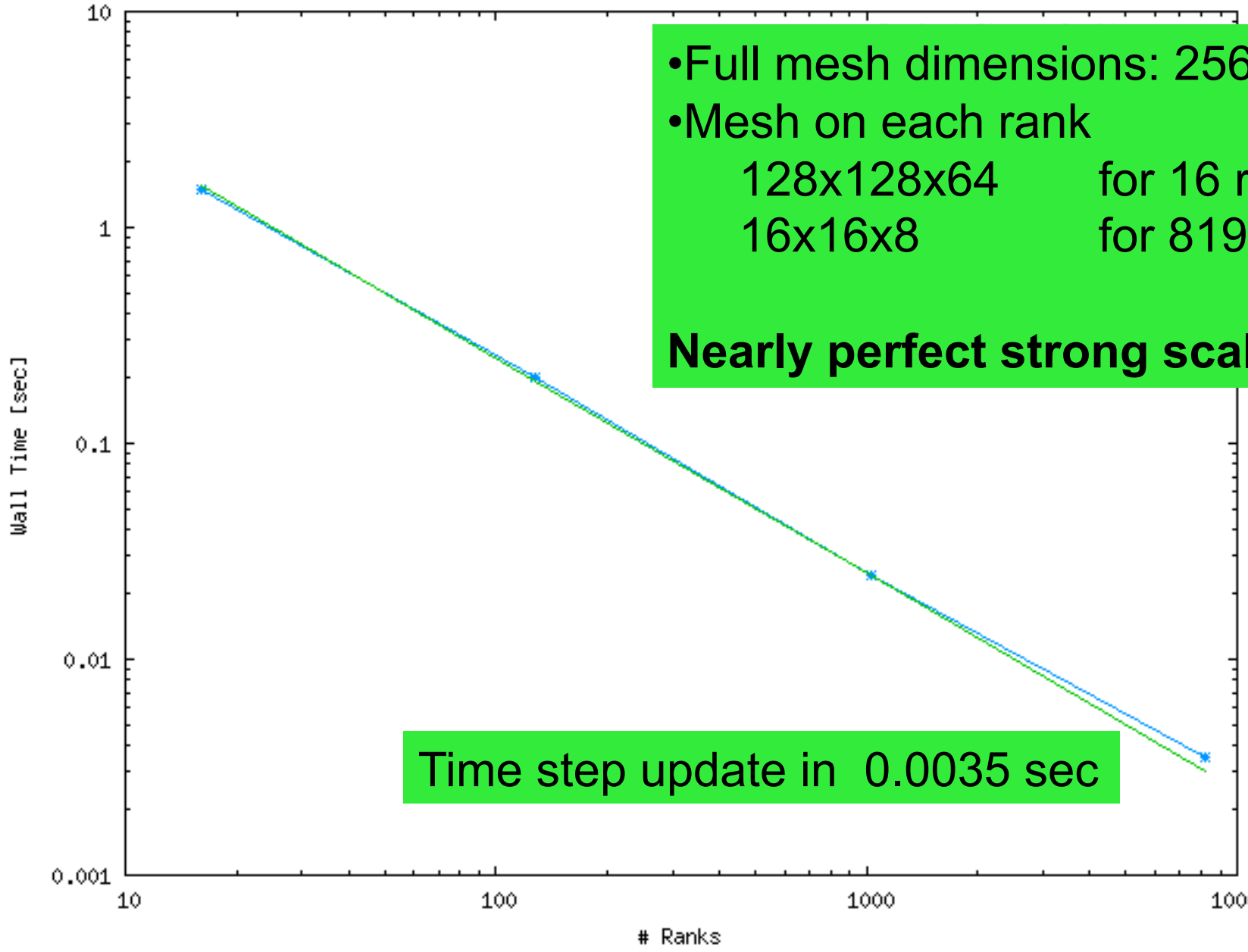
Weak Scaling



- Mesh on each rank: 128x128x64
- # of Ranks: 16, 128, 1024, 8192
- Full mesh dimensions: from 256^3 to 2048^3

Nearly perfect weak scaling

Extreme Strong Scaling



- Full mesh dimensions: 256^3
- Mesh on each rank
  - 128x128x64 for 16 ranks
  - 16x16x8 for 8192 ranks

Nearly perfect strong scaling

Time step update in 0.0035 sec

NESOTA
ver[SM]

Factors For Scaling & Performance

Within your code

- Avoid global barriers, especially mpi_alltoallv
- Use non-blocking sends & receives (mpi_isend, mpi_irecv)
- Overlap communication and computation

MPI Rank Placement

- Can shuffle host list with PMPI
- Place IO ranks on separate nodes
- Place ranks that need to communicate near each other

PBS Script & Hostfile

```
#!/bin/bash
 #PBS -l nodes=1024:ppn=8,pmem=2gb,walltime=00:20:00
 #PBS -joe
 cd <your working directory>
 mpirun –np 8192 -hostfile $PBS_NODEFILE ./a.out
```

Rank	$PBS_NODEFILE	Comments
0	node1078	Node repeated 8 times
1	Node1078	
2	Node1078	
	…	
8	node1077	Consecutive decreasing nodes, if available
	…	
16	node1076	
	…	

PBS Script & Hostfile

```
#!/bin/bash
 #PBS -l nodes=128:ppn=8,pmem=2gb,walltime=00:20:00
 #PBS -joe
 cd <your working directory>
 permute.exe  <  $PBS_NODEFILE  >  newhostfile
 mpirun –np 1024-hostfile newhostfile ./a.out
```

Script of program: “permute.exe”

- Custom for each application
- Shuffle host list to minimize communication across higher tiers of fiber fabric
- Place IO ranks on separate nodes

Using >1024 Ranks: PMPI

```
#!/bin/bash
 #PBS -l nodes=1024:ppn=8,pmem=2gb,walltime=00:20:00
 #PBS –joe

 module load pmpi/intel
 export MPI_MAX_REMSH=16
 export MPI_MAX_MPID_WAITING=64

 cd <your working directory>
permute.exe  <  $PBS_NODEFILE  >  newhostfile
 mpirun –np 8192-hostfile newhostfile ./a.out
```

Considerations For Scaling & Performance

Job scheduling vs. job size

- Be aware of IB fabric hierarchy
- Schedule jobs to fit in the hierarchy
- Avoid fragmentation of nodes

Cores	Level of Hierarchy	Average Connectivity
8	8 cores per node	Local Memory
128	16 nodes per leaf	1 QDR per node
1152	9 “near” leafs	0.5 QDR per node
8536	68 “far” leafs	0.1-0.5 QDR per node

Conclusions For Scaling & Performance

[1-128] ranks

- Should port extremely well onto Itasca
- Will typically fit within a leaf switch

[129-1152] ranks

- Should scale at least as well as on Blade or Calhoun
- Will typically fit on “near” set of 9 leaf switches

[1153-8563] ranks

- Some applications already scale extremely well to 8192
- Currently need to be aware of IB fabric hierarchy
- MSI staff is working with vendors to improve IB performance
- Dynamic routing may be needed

Thank You