

ASL Exercise 7

2K Factorial Designs &
Sections 4 and 5 Experiments

4. Throughput for Writes

Number of servers	3
Number of client machines	3
Instances of memtier per machine	2
Threads per memtier instance	1
Virtual clients per thread	[1..32]
Workload	Write-only
Multi-Get behavior	N/A
Multi-Get size	N/A
Number of middlewares	2
Worker threads per middleware	[8..64]
Repetitions	3 or more (at least 1 minute each)

- Plot throughput and response time as measured on the middleware. Combine the measurements from the two middlewares.
- The measurements should be performed for 8, 16, 32, and 64 worker threads per middleware.

4. Throughput for Writes

Summary

Maximum throughput for the full system

	WT=8	WT=16	WT=32	WT=64
Throughput (Middleware)				
Throughput (Derived from MW response time)				
Throughput (Client)				
Average time in queue				
Average length of queue				
Average time waiting for memcached				

- Based on the data provided in these tables, draw conclusions on the state of your system for a variable number of worker threads.

5. Gets and Multi-gets

5.1 Sharded

Number of servers	3
Number of client machines	3
Instances of memtier per machine	2
Threads per memtier instance	1
Virtual clients per thread	2
Workload	ratio=1:<Multi-Get size>
Multi-Get behavior	Sharded
Multi-Get size	[1..9]
Number of middlewares	2
Worker threads per middleware	max. throughput config.
Repetitions	3 or more (at least 1 minute each)

- Plot average response time as measured on the client, as well as the 25th, 50th, 75th, 90th and 99th percentiles.
- Run multi-gets with 1, 3, 6 and 9 keys (memtier configuration) with sharding enabled (multi- gets are broken up into smaller multi-gets and spread across servers).
- Provide a detailed analysis of the results (e.g., bottleneck analysis, component utilizations, average queue lengths, system saturation).

5. Gets and Multi-gets:

5.2 Non-Sharded

Number of servers	3
Number of client machines	3
Instances of memtier per machine	2
Threads per memtier instance	1
Virtual clients per thread	2
Workload	ratio=1:<Multi-Get size>
Multi-Get behavior	Non-Sharded
Multi-Get size	[1..9]
Number of middlewares	2
Worker threads per middleware	max. throughput config.
Repetitions	3 or more (at least 1 minute each)

- Plot average response time as measured on the client, as well as the 25th, 50th, 75th, 90th and 99th percentiles.
- Run multi-gets with 1, 3, 6 and 9 keys (memtier configuration) with sharding disabled.
- Provide a detailed analysis of the results (e.g., bottleneck analysis, component utilizations, average queue lengths, system saturation).

5. Gets and Multi-gets:

5.3 Histogram

For the case with 6 keys inside the multi-get, display four histograms representing response time distribution as measured in the following configurations:

- Sharded as measured on the MW.
- Sharded as measured on the Clients,
- Non-sharded as measured on the MW.
- Non-sharded as measured on the Clients.

Choose the bucket size in the same way for all four, and such that there are at least 10 buckets on each of the graphs.

5. Gets and Multi-gets:

5.4 Summary

- Provide a detailed comparison of the sharded and non-sharded modes.
- For which multi-GET size is sharding the preferred option?
- Provide a detailed analysis of your system.
- Add any additional figures and experiments that help you illustrate your point and support your claims.

2K Factorial Designs

How do you determine the effect of k factors on a system?

- Design experiments, where each factor can have 2 values.
- Perform 2^k experiments.
- Analyze the results, giving insights on:
 1. The effect of factors alone on the result.
 2. The contributed effect of factor together on the result.

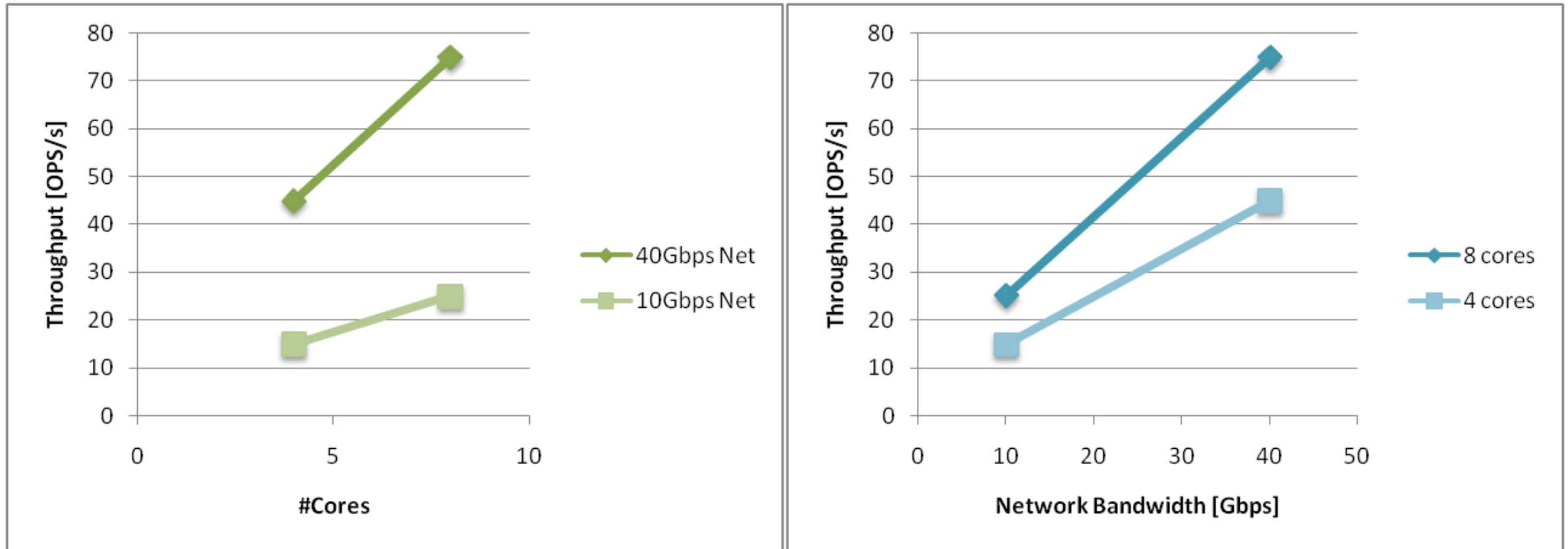
Example:

Factor A: Network Bandwidth Values: 10 Gbit/s, 40 Gbit/s

Factor B: # CPU Cores Values: 4, 8

Measure throughput (requests/second)

What if we plot it?



- Difficult to look at two graphs at the same time
- Hard to reach definitive conclusion

Use a table-based representation instead:

Factor A: Network Bandwidth Values: 10 Gbit/s, 40 Gbit/s

Factor B: # CPU Cores Values: 4, 8

Measure throughput (requests/second)

Define 2 variables: x_A , x_B

x_A $\begin{cases} -1 & \text{if 10 Gbit/s} \\ 1 & \text{if 40 Gbit/s} \end{cases}$

x_B $\begin{cases} -1 & \text{if 4 Cores} \\ 1 & \text{if 8 Cores} \end{cases}$

Example:

Experiment	xA (bandwidth)	xB (# cores)	Y (TPUT)
1	-1 (10 Gbit/s)	-1 (4 cores)	15
2	1 (40 Gbit/s)	-1 (4 cores)	45
3	-1 (10 Gbit/s)	1 (8 cores)	25
4	1 (40 Gbit/s)	1 (8 cores)	75

Example:

Define 2 variables: x_A , x_B

$$x_A \begin{cases} -1 & \text{if 10 Gbit/s} \\ 1 & \text{if 40 Gbit/s} \end{cases}$$

$$x_B \begin{cases} -1 & \text{if 4 Cores} \\ 1 & \text{if 8 Cores} \end{cases}$$

Solve the equation:

$$\text{throughput } y = q_0 + \text{Effect of bandwidth } q_A * x_A + \text{Effect of \# Cores } q_B * x_B + \text{Combined Effect } q_{AB} * x_A * x_B$$

Example:

Experiment	xA (bandwidth)	xB(# cores)	y
1	-1 (10 Gbit/s)	-1 (4 cores)	y1
2	1 (40 Gbit/s)	-1 (4 cores)	y2
3	-1 (10 Gbit/s)	1 (8 cores)	y3
4	1 (40 Gbit/s)	1 (8 cores)	y4

Measured
Values

$$\text{throughput } y = q_0 + \text{Effect of bandwidth } q_A * x_A + \text{Effect of \# Cores } q_B * x_B + \text{Combined Effect } q_{AB} * x_A * x_B$$

Example:

I	x _A	x _B	x _A *x _B	y
1	-1	-1	1	15
1	1	-1	-1	45
1	-1	1	-1	25
1	1	1	1	75
40	20	10	5	Total/4

So-called
Effects!


$$y = q_0 + q_A * x_A + q_B * x_B + q_{AB} * x_A * x_B$$

Allocation of Variation

- Importance of a factor: Proportion of the total variation in the result that is explained by the factor.
- Total variation of $y = SST = \sum_{i=1}^{2^2} (y_i - \bar{y})^2$
- For a 2^2 design, $SST = 2^2(qA^2 + qB^2 + qAB^2)$



Variation explained by A

Variation explained by B

Variation explained by
The interaction of A and B

Example showing the allocation of variation

Experiment	xA (bandwidth)	xB(# cores)	Y (Throughput)
1	-1 (10 Gbit/s)	-1 (4 cores)	15
2	1 (40 Gbit/s)	-1 (4 cores)	45
3	-1 (10 Gbit/s)	1 (8 cores)	25
4	1 (40 Gbit/s)	1 (8 cores)	75

$$SST = 2^2(20^2 + 10^2 + 5^2) = 2100$$

$$SSA = 2^2(20^2) = 1600$$

$$SSB = 2^2(10^2) = 400$$

$$SSAB = 2^2(5^2) = 100$$

SSA/SST

SSB/SST

SSAB/SST

Parameter	Mean Estimate	Variation Explained (%)
q0	40	
qA	20	76 %
qB	10	19 %
qAB	5	5 %

Interpretation

Parameter	Mean Estimate	Variation Explained (%)
q0	40	
qA	20	76 %
qB	10	19 %
qAB	5	5 %

- Average throughput is 40 reqs/second
- The throughput is mostly affected by the network bandwidth: 76 % of the variation
- The number of cores contributes 19 % to the variation
- The interactive effect of network bandwidth and number of cores is relatively less: only 5 %

How about replication: $2^k r$ Factorial Designs

Isolate experimental errors!

Solve the equation:

$$y = q_0 + q_A * x_A + q_B * x_B + q_{AB} * x_A * x_B + \textcircled{e}$$

Experimental error!

Example:

						$y - y_{\text{mean}}$			
I	xA	xB	xA*xB	y	y_mean	i	e1	e2	e3
1	-1	-1	1	(15,18,12)	15	1	0	3	-3
1	1	-1	-1	(45,48,51)	48	2	-3	0	3
1	-1	1	-1	(25,28,19)	24	3	1	4	-5
1	1	1	1	(75,75,81)	77	4	-2	-2	4
41	21.5	9.5	5	Total/4					


Sum of squared errors:

$$\text{SSE} = \sum_{i=1}^{2^k} \sum_{j=1}^r e_{ij}^2$$

SSE = 102

Allocation of variation:

$$SST = SSA + SSB + SSAB + SSE$$


$$SST = 2^2 r \cdot qA^2 + 2^2 r \cdot qB^2 + 2^2 r \cdot qAB^2 + \sum_i^{2^2} \sum_j^r e_{ij}^2$$

What if the effects are multiplicative?

- Until now we have assumed the model:

$$y = q_0 + \boxed{q_A x_A} + \boxed{q_B x_B} + \boxed{q_{AB} x_A x_B} + \boxed{e}$$

All effects are added!

- If the effects multiply (for example, A: size of a workload and B: clock frequency)
- Compute in log domain!

$$y = A * B$$



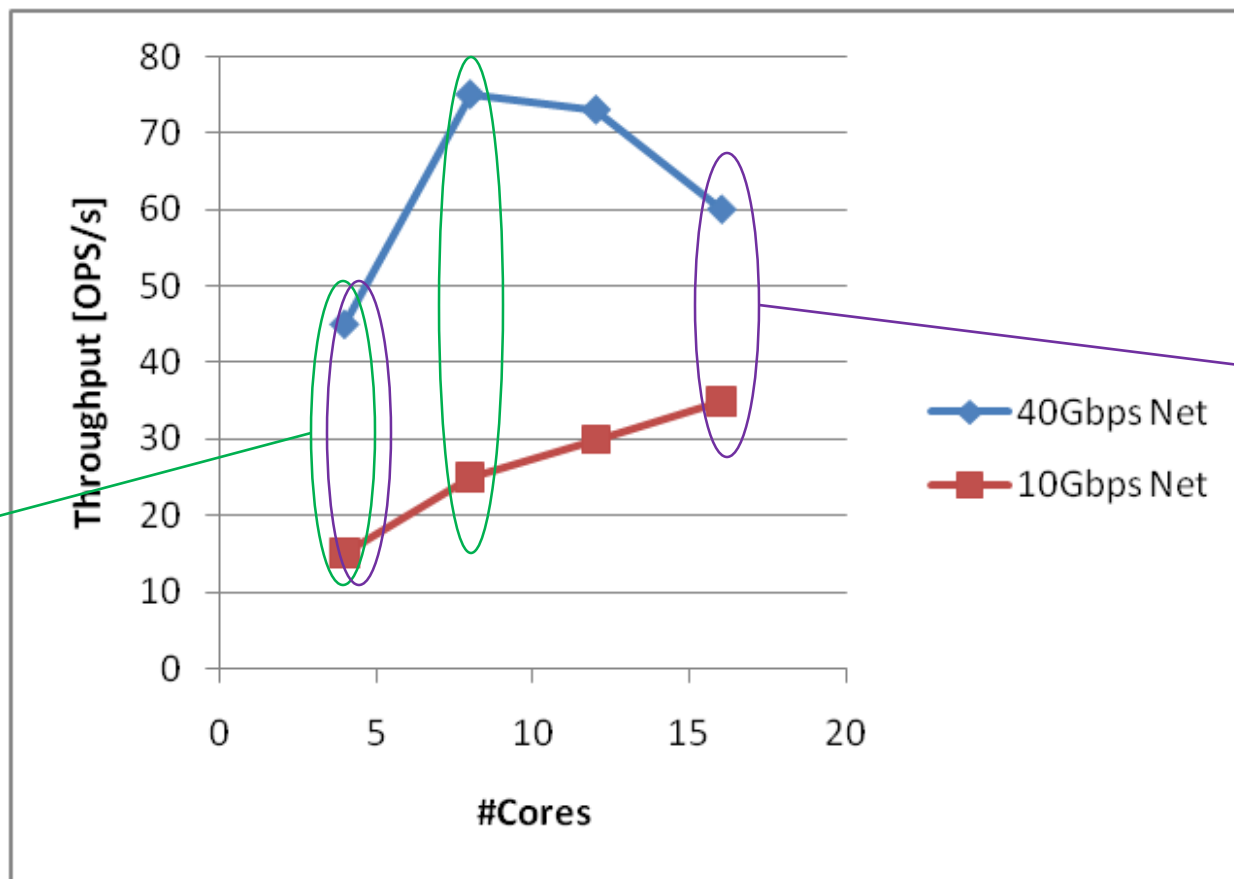
$$\log(y) = \log(A) + \log(B)$$

- After taking the log of the measurements, you can use the model exactly as before.

Discussion

- What happens if we do a 2k on the following graph? Which points to choose?

Conclusion could be that choice of network has much higher effect



Conclusion could be that both have equal effect