Tradeoffs in Designing Web Clusters

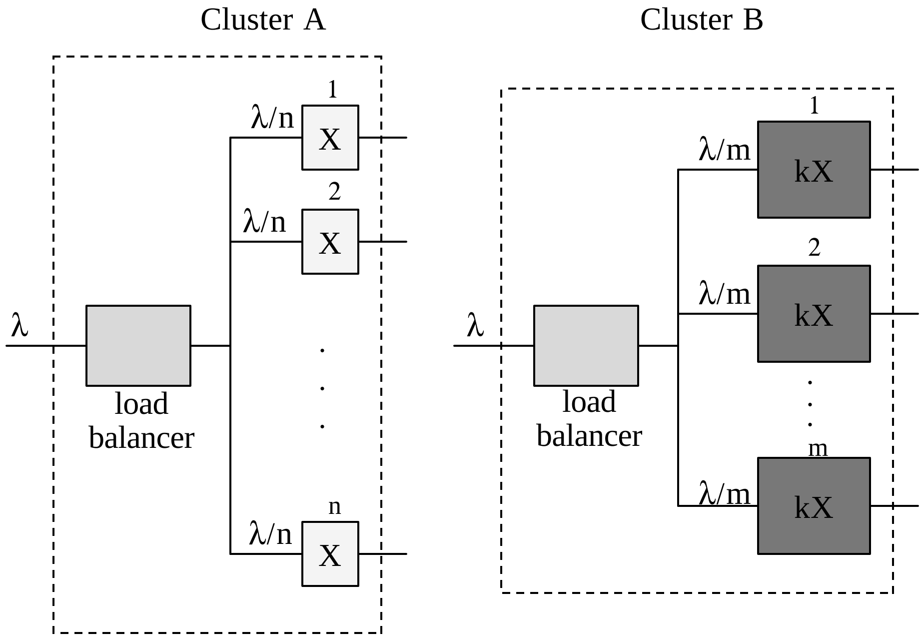
Daniel A. Menascé
Department of Computer Science
George Mason University

1

Typical Questions

- Should I use a large number of low-capacity inexpensive servers or a small number of high-capacity costly ones?
- How many servers of a given type are required to provide a certain performance level at a given cost?
- How many servers are needed to build a Web site with a given reliability?

2



Comparison Criteria

- Equal Average Response Time
- Equal Cluster Capacity: nX = mkX.
  - m = n / k.
- Equal Cost: n C(X) = m C(kX).
  - m = n C(x) / C(kX)
- Equal Reliability.

Reliability Considerations

$$R_A = 1 - (1 - r_A)^n$$

$$R_B = 1 - (1 - r_B)^m$$

$$R_A = R_B \quad \Rightarrow$$

$$m = \frac{n \log(1 - r_A)}{\log(1 - r_B)}$$

Basic Performance Model (M/G/1)

$$T = S + \frac{U \times S(1 + C^2)}{2(1 - U)}$$

- T: average response of a Web request
- S: average request service time
- C: coefficient of variation of the service time
- U: server utilization, equal to $l_w S$

The utilization has to be < 1 …

$$U_A = \frac{l}{n} \times \frac{1}{X} < 1 \quad \Rightarrow l < nX$$

$$U_B = \frac{l}{m} \times \frac{1}{kX} < 1 \quad \Rightarrow l < mkX$$

7

Response Time Equations

$$T_A = \frac{1}{X} + \frac{\frac{l}{n} \times \left(\frac{1}{X}\right)^2 \left(1 + C^2\right)}{2\left(1 - \frac{l}{n} \times \frac{1}{X}\right)}$$

$$T_B = \frac{1}{kX} + \frac{\frac{l}{m} \times \left(\frac{1}{kX}\right)^2 \left(1 + C^2\right)}{2\left(1 - \frac{l}{m} \times \frac{1}{kX}\right)}$$

8

Average Number of Requests

- From Little’s Law:

$$N_A = l T_A$$

$$N_B = l T_B$$

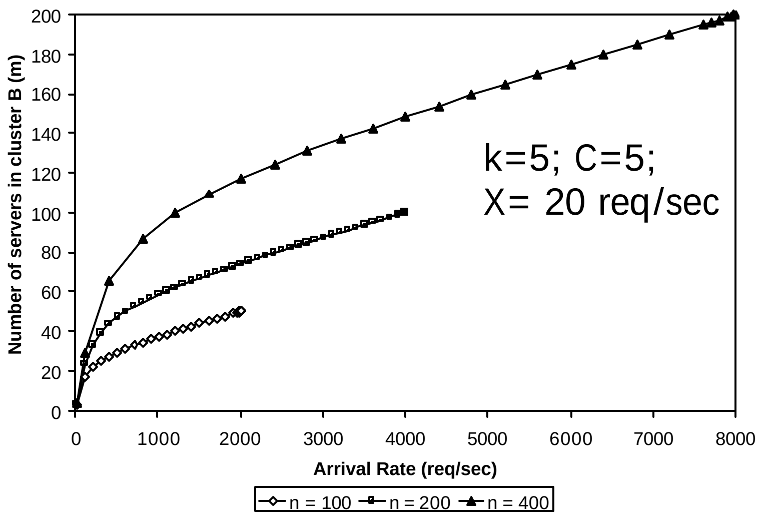
Equal Response Time Case

$$T_A = T_B \quad \Rightarrow$$

$$m = \frac{1}{kX}\left[l + \frac{1}{\frac{2(k-1)}{l(1+C^2)} + \frac{k}{nX - l}} \right]$$

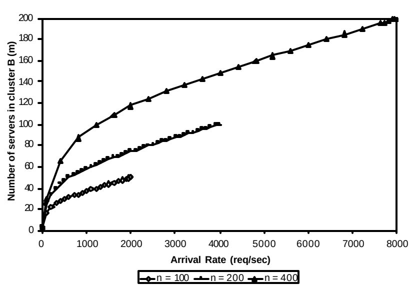
$$\lim_{l \to nX} m = n / k$$

Equal Response Time Case



11

Equal Response Time Case



- When cluster A has 400 servers and λ = 4,810 req/sec, c cluster B needs 160 servers to obtain the same avg. response time of 1.03 sec.

- For a sufficiently large value of λ, m increases linearly with λ at the rate of $(k-1)/(k^2X)$

12

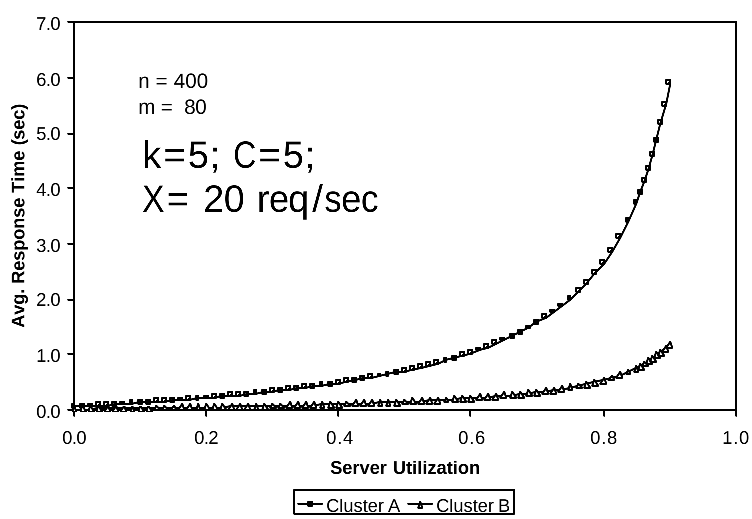
Equal Capacity Case

- m = n/k
- Thus, $T_B = T_A / k$ $\qquad N_B = N_A / k$

- Cluster B average response time is always k times less than that of cluster A.
- Cluster B can handle 1/k of the requests that cluster A can handle.

13

Equal Capacity Case



14

Equal Cost Case

- If the cost is proportional to the capacity, then m has the same value as in the equal capacity case.
- Sublinear cost function: there is some economy of scale (i.e., the cost per unit of capacity decreases with the capacity).
- Superlinear cost function: the server cost per unit of capacity increases with the capacity.

15

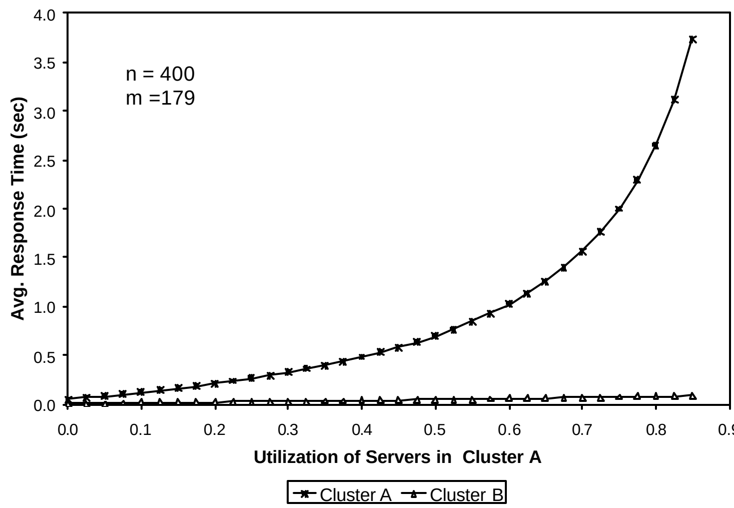
Equal Cost Case: sub-linear cost function

$$C(x) = a\sqrt{x} \Rightarrow m = m / \sqrt{k}$$

$$U_A = U_B \sqrt{k}$$

16

Equal Cost Case: sublinear cost function



17

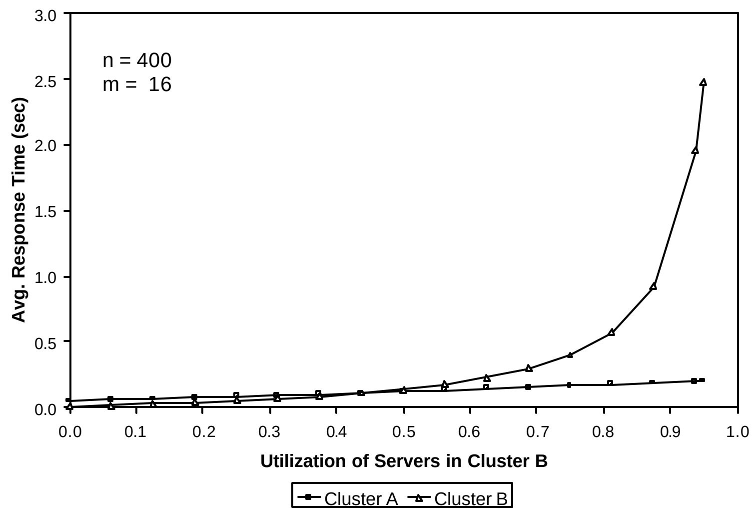
Equal Cost Case: superlinear cost function

$$C(x) = ax^2 \Rightarrow m = m / k^2$$

$$U_B = kU_A$$

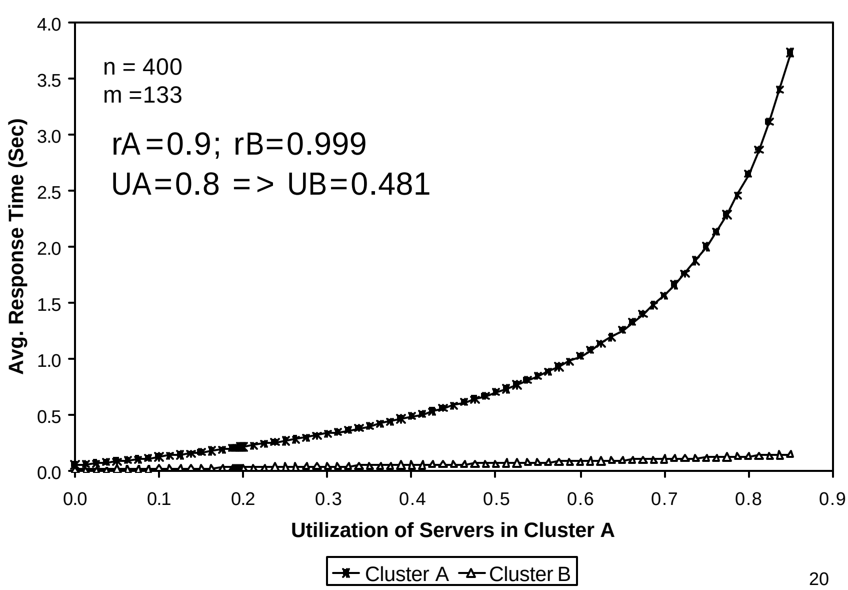
18

Equal Cost Case: superlinear cost function



19

Equal Reliability Case



20

Comparison: λ = 4,800 req/sec; TA= 1.025 sec; UA= 0.6; n = 400

T_B (sec)	m	U_B	Case
1.025	54	89%	Equal response time.
0.205	80	60%	Equal total capacity
0.058	179	27%	Equal cost. $C(x)=vx$
8	16	> 100%	Equal cost. $C(x) = x^2$
0.083	133	36%	Equal cluster reliability.

21