Economies of Scale in Parallel-Server Systems

Josu Doncel[1]
Inria (France)

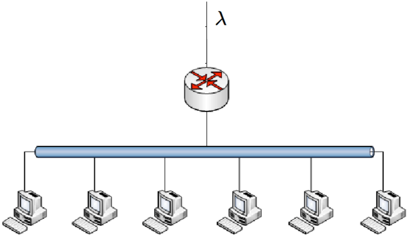
joint work with S. Aalto (Aalto University) and U. Ayesta (IRIT-CNRS, Ikerbasque and UPV/EHU)

IEEE Infocom 2017
Atlanta, USA

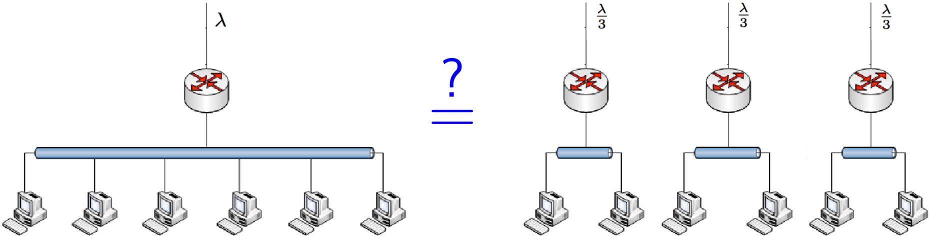
May 3, 2017

[1] Currently at University of the Basque Country (Spain)

Load Balancing of Parallel Server Systems



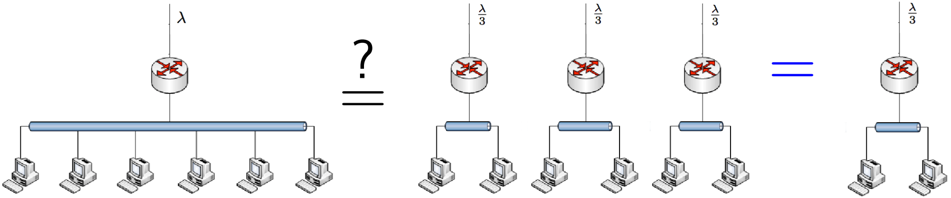
Load Balancing of Parallel Server Systems



Decentralization

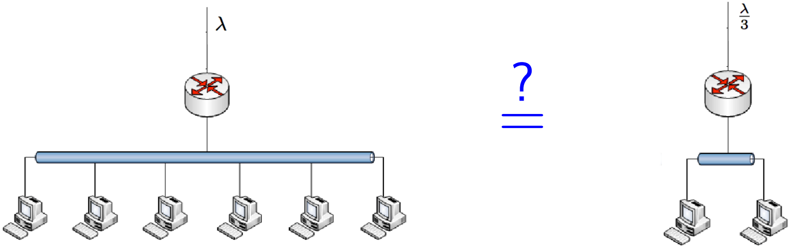
Performance Degradation

Load Balancing of Parallel Server Systems



Symmetric $\Rightarrow$ Equal performance!

Load Balancing of Parallel Server Systems

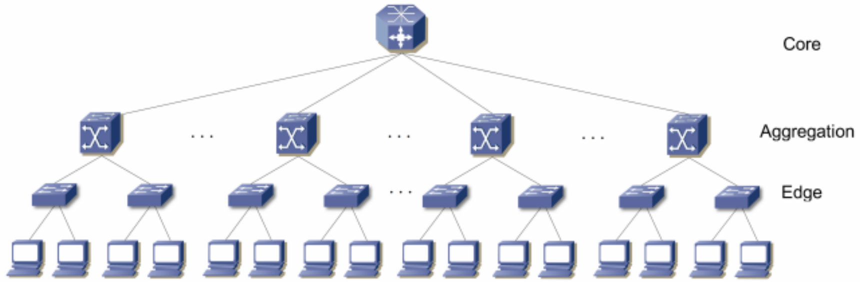


$\stackrel{?}{=}$

Economies of Scale

Arrival rate and number of servers

Application



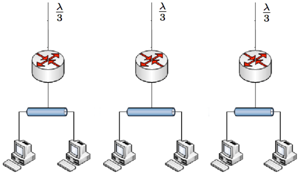
Outline

Outline

Degradation Factor

$\mathbb{E}(W(K, n, x_m, x_M, \lambda))$

- K: FCFS homogeneous servers
- n: number of groups
- x_m: minimum job size
- x_M: maximum job size
- λ: arrival rate (Poisson)

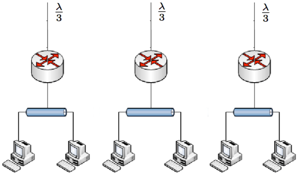


K=6, n=3

Degradation Factor

$\mathbb{E}(W(K, n, x_m, x_M, \lambda))$

- K: FCFS homogeneous servers
- n: number of groups
- x_m: minimum job size
- x_M: maximum job size
- λ: arrival rate (Poisson)



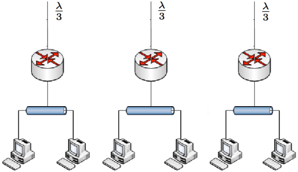
K=6, n=3

$$\mathbb{E}(W(K, n, x_m, x_M, \lambda)) = \mathbb{E}\left(W\left(\frac{K}{n}, 1, x_m, x_M, \frac{\lambda}{n}\right)\right)$$

Degradation Factor

$\mathbb{E}(W(K, n, x_m, x_M, \lambda))$

- K: FCFS homogeneous servers
- n: number of groups
- x_m: minimum job size
- x_M: maximum job size
- λ: arrival rate (Poisson)



K=6, n=3

$$\mathbb{E}(W(K, n, x_m, x_M, \lambda)) = \mathbb{E}\left(W\left(\frac{K}{n}, 1, x_m, x_M, \frac{\lambda}{n}\right)\right)$$

Definition (Degradation Factor)

$$D(K, n, x_m, x_M, \lambda) = \frac{\mathbb{E}\left(W\left(\frac{K}{n}, 1, x_m, x_M, \frac{\lambda}{n}\right)\right)}{\mathbb{E}\left(W\left(K, 1, x_m, x_M, \lambda\right)\right)}$$

Degradation Factor (cont.)

$$D(K, n, x_m, x_M, \lambda) = \frac{\mathbb{E}\left(W\left(\frac{K}{n}, 1, x_m, x_M, \frac{\lambda}{n}\right)\right)}{\mathbb{E}\left(W\left(K, 1, x_m, x_M, \lambda\right)\right)}$$

$\Rightarrow \mathbb{E}\left(W\left(R, 1, x_m, x_M, \bar{\lambda}\right)\right)$

- $R = K/n$ and $\bar{\lambda} = \lambda/n$
- $R = K$ and $\bar{\lambda} = \lambda$

Degradation Factor (cont.)

$$D(K,n,x_m,x_M,\lambda) = \frac{\mathbb{E}\left(W\left(\frac{K}{n},1,x_m,x_M,\frac{\lambda}{n}\right)\right)}{\mathbb{E}\left(W\left(K,1,x_m,x_M,\lambda\right)\right)}$$

$\Rightarrow \mathbb{E}\left(W\left(R,1,x_m,x_M,\bar{\lambda}\right)\right)$

- $R = K/n$ and $\bar{\lambda} = \lambda/n$
- $R = K$ and $\bar{\lambda} = \lambda$

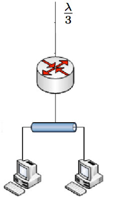
Case: n=K

- Queue with arrival rate λ/K

SITA-E Scheduling

Cut-offs: $x_0, x_1, \ldots, x_{K-1}, x_K$
$(x_m = x_0, x_M = x_K)$

- Server i: $[x_{i-1}, x_i]$
- Equal load

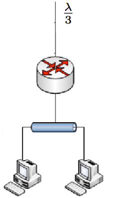


$$\int_{x_0=x_m}^{x_1} xf(x)dx = \int_{x_1}^{x_2} xf(x)dx = \cdots = \int_{x_{K-1}}^{x_K=x_M} xf(x)dx.$$

SITA-E Scheduling

Cut-offs: $x_0, x_1, \ldots, x_{K-1}, x_K$
($x_m = x_0, x_M = x_K$)

- Server i: $[x_{i-1}, x_i]$
- Equal load



$$\int_{x_0=x_m}^{x_1} xf(x)dx = \int_{x_1}^{x_2} xf(x)dx = \cdots = \int_{x_{K-1}}^{x_K=x_M} xf(x)dx.$$

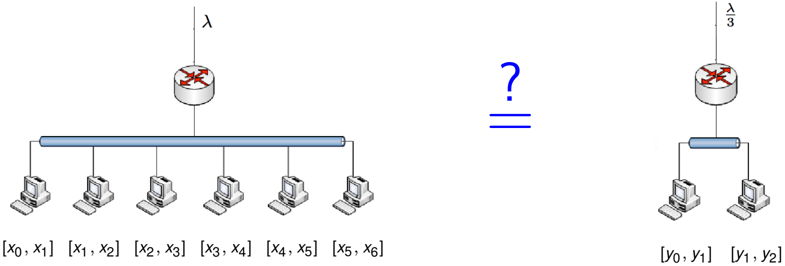
Disadvantages

- Not optimal
  JSQ, Po2, SITA Optimal...
  $\Rightarrow$ Difficult

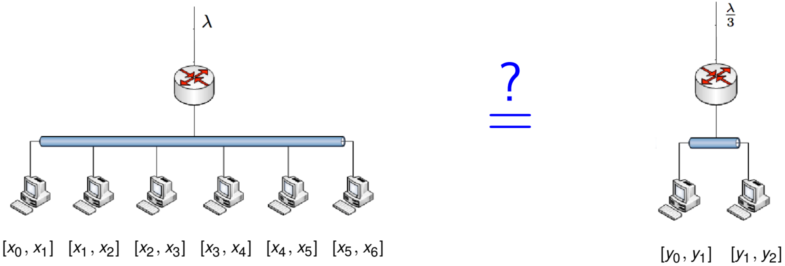
Advantages

- No signaling
- Easy implementation
- Cut-offs expression

Thresholds in SITA-E



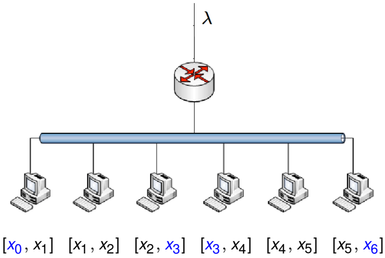
Thresholds in SITA-E



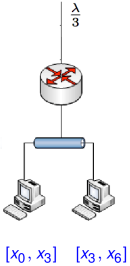
Lemma

If $f(x) > 0$, then $x_{i \cdot n} = y_i$, $i = 0, \ldots, K/n$

Thresholds in SITA-E



$\stackrel{?}{=}$



Lemma

If $f(x) > 0$, then $x_{i \cdot n} = y_i$, $i = 0, \ldots, K/n$

$\Rightarrow$ Only required: $x_0, \ldots, x_K$

Influence of x_m and x_M

$\gamma = \frac{x_m}{x_M} \in [0, 1]$

Lemma

If $\gamma = 1$, then $D(K, n, x_m, x_M, \lambda) = 1$.

$\Rightarrow$ Deterministic

Influence of x_m and x_M

$\gamma = \frac{x_m}{x_M} \in [0, 1]$

Lemma

If $\gamma = 1$, then $D(K, n, x_m, x_M, \lambda) = 1$.

$\Rightarrow$ Deterministic

If the degradation decreases with γ

$$\lim_{\gamma \to 1} D(K, n, x_m, x_M, \lambda) \leq D(K, n, x_m, x_M, \lambda) \leq \lim_{\gamma \to 0} D(K, n, x_m, x_M, \lambda)$$

Question: Is it always true?

Outline

Uniform Distribution

$$f(x) = \frac{1}{x_M - x_m}, \; x_m \leq x \leq x_M$$

Two servers

$$1 \leq \mathrm{D}(K, n, x_m, x_M, \lambda) \leq 1.138.$$

$K > 2$ servers

Assume that the degradation decreases with γ,

$$1 \leq \mathrm{D}(K, n, x_m, x_M, \lambda) \leq 4/3.$$

$\Rightarrow$ Small: Higher variability?

Bounded Pareto

$$f(x) = \frac{\alpha \, x_m^{\alpha}}{1 - (x_m/x_M)^{\alpha}} \, x^{-\alpha-1}, \; x_m \leq x \leq x_M$$

Case $\alpha = 1$

$$1 \leq \mathrm{D}(K, n, x_m, x_M, \lambda) \leq \infty$$

Case $\alpha \neq 1$

Assume the degradation decreases with γ

$$1 \leq \mathrm{D}(K, n, x_m, x_M, \lambda) \leq n^{\frac{1}{|1-\alpha|}}.$$

$\Rightarrow$ Increases with variability of jobs.

Two Points

$$f(x) = \begin{cases} p, & \text{if } x = x_m, \\ 1-p, & \text{if } x = x_M. \end{cases}$$

Maximizes variance (bounded and fixed support)

Two Points

$$f(x) = \begin{cases} p, & \text{if } x = x_m, \\ 1-p, & \text{if } x = x_M. \end{cases}$$

Maximizes variance (bounded and fixed support)

2 servers and equal load

$$1 \leq \mathrm{D}(K, n, x_m, x_M, \lambda) \leq \infty$$

2 servers and unequal load

$$1 \leq \mathrm{D}(K, n, x_m, x_M, \lambda) \leq \infty$$

Outline

Degenerate Hyper-exponential

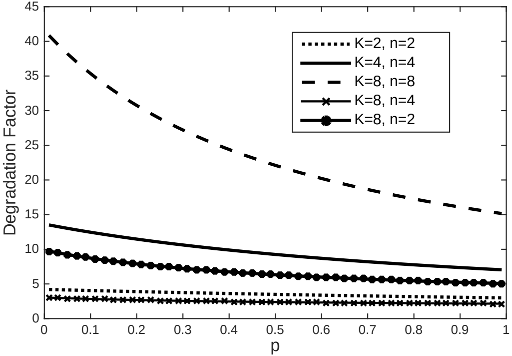
$$\text{Exponential} \begin{cases} \text{rate } \mu p, & \text{w.p. } p, \\ \text{rate } \infty, & \text{w.p. } 1-p, \end{cases}$$

- Mean: $1/\mu$
- Variance: $\frac{1}{p\mu^2}$

Degenerate Hyper-exponential

$$\text{Exponential} \begin{cases} \text{rate } \mu p, & \text{w.p. } p, \\ \text{rate } \infty, & \text{w.p. } 1-p, \end{cases}$$

- Mean: $1/\mu$
- Variance: $\frac{1}{p\mu^2}$



SITA Optimal Thresholds

[Harchol and Vesilo, 2010]

- Mean response time: unknown (two servers)
- 2 servers and $\gamma = 9/10^{14}$

SITA Optimal Thresholds

[Harchol and Vesilo, 2010]

- Mean response time: unknown (two servers)
- 2 servers and $\gamma = 9/10^{14}$

	Optimal SITA Degradation Factor		
	$\rho = 0.005$	$\rho = 0.5$	$\rho = 0.8$
$\alpha = 0.25$	333.74	87.77	8.6594
$\alpha = 0.5$	$2.2476 \cdot 10^4$	4219.9	18.7679
$\alpha = 0.75$	$3.3604 \cdot 10^5$	$1.3187 \cdot 10^5$	133.8889
$\alpha = 1.25$	$3.3604 \cdot 10^5$	$1.3187 \cdot 10^5$	133.8889
$\alpha = 1.5$	$2.2476 \cdot 10^4$	4219.9	18.7679
$\alpha = 1.75$	333.74	87.77	8.6594

Outline

Conclusions

- SITA-E
- FCFS homogeneous servers
- Particular distributions: Uniform, Bounded Pareto and Two Points

Conclusion

Scaling $\Rightarrow$ non-negligible degradation

- Variability of jobs is high

Future Work

- SITA-E
- FCFS homogeneous servers
- Particular distributions: Uniform, Bounded Pareto and Two Points

Future Work

- SITA-E

  $\Rightarrow$ JSQ, Po2, SITA Optimal...
- FCFS homogeneous servers
- Particular distributions: Uniform, Bounded Pareto and Two Points

Future Work

- SITA-E

  $\Rightarrow$ JSQ, Po2, SITA Optimal...
- FCFS homogeneous servers

  $\Rightarrow$ Heterogeneous servers? PS queues?
- Particular distributions: Uniform, Bounded Pareto and Two Points

Future Work

- SITA-E

  $\Rightarrow$ JSQ, Po2, SITA Optimal...
- FCFS homogeneous servers

  $\Rightarrow$ Heterogeneous servers? PS queues?
- Particular distributions: Uniform, Bounded Pareto and Two Points

  $\Rightarrow$ General distribution? Lower-bounded by 1, monotonicity with γ

Future Work

- SITA-E

  $\Rightarrow$ JSQ, Po2, SITA Optimal...
- FCFS homogeneous servers

  $\Rightarrow$ Heterogeneous servers? PS queues?
- Particular distributions: Uniform, Bounded Pareto and Two Points

  $\Rightarrow$ General distribution? Lower-bounded by 1, monotonicity with γ

Other performance measures:

- Tail-probabilities?
- Second moment of waiting time?

Thank you very much

Questions?