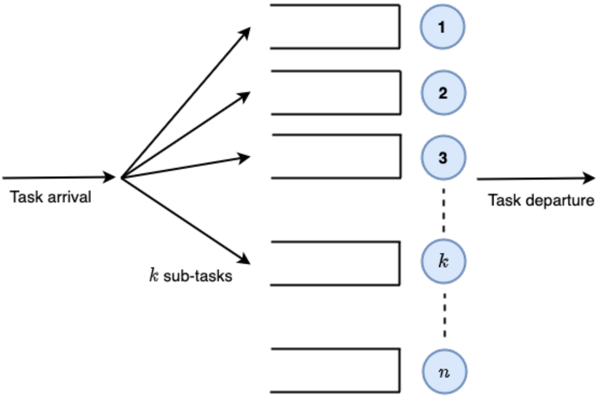
Fork-join scheduling on heterogeneous parallel servers

Parimal Parag
Moonmoon Mohanty
Gaurav Gautam
Rooji Jinan
Vaneet Aggarwal

Electrical Communication Engineering
Indian Institute of Science

(k, k) fork-join system



(k, k) fork-join system

The task departs from the system on completion of all k sub-tasks

Distributed Computing

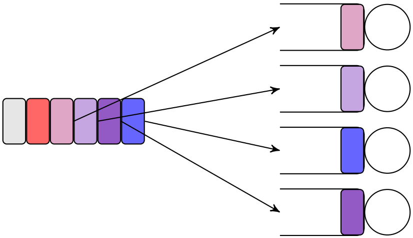
Importance

This system is a critical building block in the job processing workflow of many data center services including web search, social networking, and big data analytics, that constitutes a major part of job processing time and hardware cost. For example, MapReduce used in large-scale data processing frameworks.

Challenges

- Heterogeneous environment - slow and fast servers
- Optimization of task completion time is crucial to maximize efficiency and resource utilization.

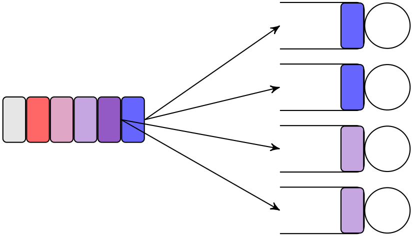
Load balancing



Splitting workloads

- Objectives: reduce completion time and energy consumption
- Observations: queue lengths, arrival and service statistics

Redundant computing



- Performance improvement due to parallel processing

Related Works

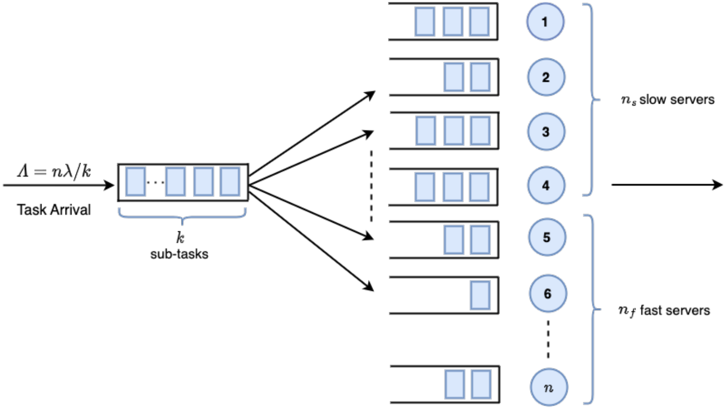
Load balancing strategies in homogeneous system

- **Without sub-division of tasks** [M. Mitzenmacher et al., 2001], [U.Ayesta et al., 2019], power-of-d variants
- **With sub-division of tasks** [A.Badita et al., 2019] [R.Jinan et al., 2022],

Load balancing strategies in heterogenous system

- **Without sub-division of tasks**[Der Boor et al., 2021], [Jaleel et al., 2022]: power-of-d variants
- **With sub-division of tasks:** Our work

Probabilistic Scheduling



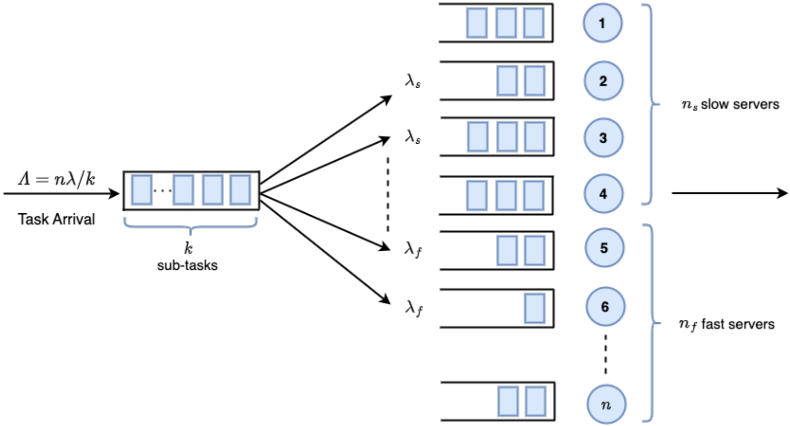
Objective

Find the optimal selection probability that minimizes the mean task completion time

Key contributions

- Establishing asymptotic independence of the stationary workload distribution in a heterogeneous server system with two classes of heterogeneity
- Analytical computation of the limiting mean task completion time for systems with an arbitrarily large number of servers
- Providing an upper bound on the limiting mean response time and identification of the probability selection parameter p_s that minimizes this bound

Parallel Compute with Heterogeneous Service



System Parameters

Random number of slow and fast servers

The probability of selecting k_s slow and $k - k_s$ fast servers for any task

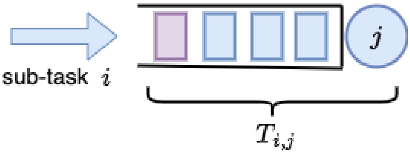
$$q(k_s) = \binom{k}{k_s} p_s^{k_s} (1 - p_s)^{k-k_s}.$$

Arrival rate at individual servers

The arrival at each server is a thinned Poisson process with the arrival rate at slow and fast servers are

$$\lambda_s \triangleq \frac{\lambda n}{k}\left(\frac{kp_s}{nf_s}\right) = \frac{\lambda p_s}{f_s}, \qquad \lambda_f \triangleq \frac{\lambda \bar{p}_s}{\bar{f}_s}.$$

Performance metrics

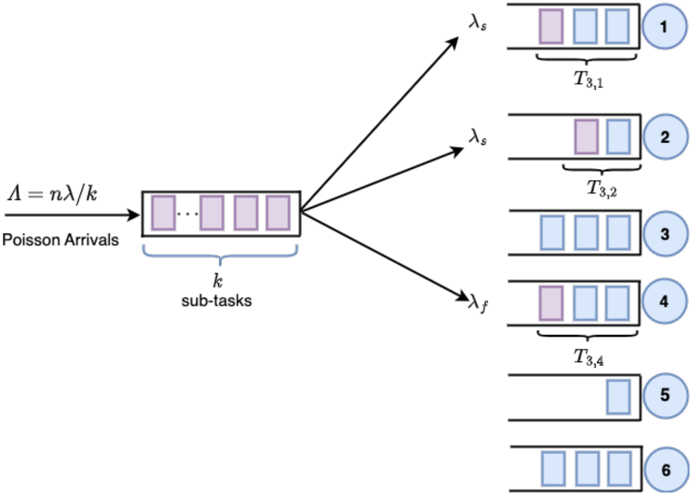


Sub-task completion time

The sub-task completion time at server j is denoted by

$$T_{i,j} \triangleq W_{i,j} + X_{i,j}$$

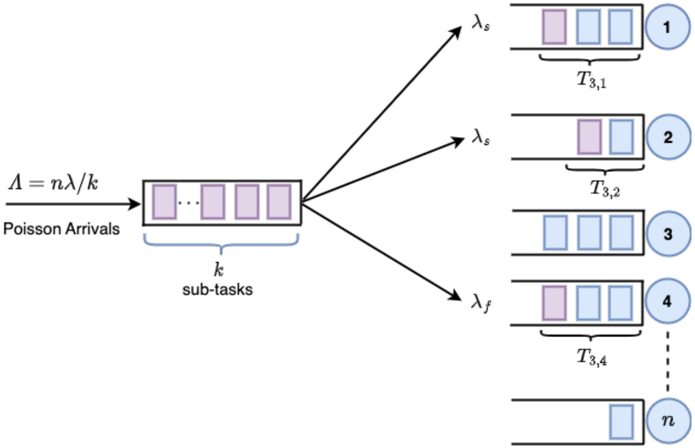
Performance metrics



Task completion time

$T_i \triangleq \max_{j \in I^i} T_{i,j}$

Asymptotic Independence



Asymptotic independence of workload at any k queues for a large number of servers

Asymptotic Independence

Theorem

If $\pi^k, \hat{\pi}^k$ are the equilibrium distributions for workloads in the first $k = o(n^{\frac{1}{4}})$ servers of systems $\mathcal{S}$ and $\hat{\mathcal{S}}$ respectively, the total variation distance

$$\lim_{n \to \infty} d_{\mathrm{TV}}(\pi^k, \hat{\pi}^k) = 0$$

Asymptotic Independence

When $k(n) = \left[n^{\frac{2}{3}}\right]$

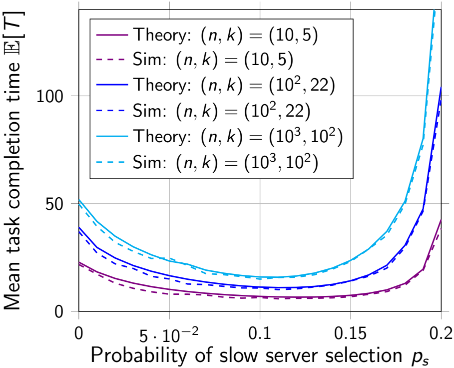


Figure: System with $f_s = 0.5$, $\lambda = 1.2$, $(\mu_s, \mu_f) = (0.5, 2.5)$.

Mean task completion time

- Recall that $T_i = \max_{j \in I^i} T_{i,j}$
- Under asymptotic independence at stationarity, we have

$$P\{T_i \leqslant x\} = \prod_{j \in I^i} P\{T_{i,j} \leqslant x\}$$

- Identical sub-task completion time distributions at stationarity: L_s for slow and L_f for fast servers

$$\mathbb{E}[T] = \sum_{k_s=0}^{k} \binom{k}{k_s} p_s^{k_s} (1-p_s)^{k_s} L_s(x)^{k_s} L_f(x)^{k-k_s}$$

Mean task completion time

Under asymptotic independence, we have

$$\mathbb{E}[T] = \int_{x \in \mathbb{R}_+} [1 - (p_s L_s(x) + \bar{p}_s L_f(x))^k] dx$$

Bound on Mean Task Completion Time

Upper and Lower Bound

- Average smaller than maximum smaller than sum, i.e.

$$\frac{1}{|I^i|}\sum_{j\in I^i} T_{i,j} \leqslant \max_{j\in I^i} T_{i,j} \leqslant \sum_{j\in I^i} T_{i,j}$$

- Mean of sum of sub-task completion times

$$\lim_{i\to\infty} \mathbb{E}\sum_{j\in I^i} T_{i,j} = kp_s \int_{x\in\mathbb{R}_+} \bar{L}_s(x)dx + k\bar{p}_s \int_{x\in\mathbb{R}_+} \bar{L}_f(x)dx$$

- For general service times

$$\lim_{i\to\infty} \mathbb{E}\sum_{j\in I^i} T_{i,j} = \frac{kp_s}{\lambda_s}\left(\rho_s + \frac{\lambda_s^2\mathbb{E}X_s^2}{2(1-\rho_s)}\right) + \frac{k\bar{p}_s}{\lambda_f}\left(\rho_f + \frac{\lambda_f^2\mathbb{E}X_f^2}{2(1-\rho_f)}\right),$$

where the load $\rho_s = \lambda_s\mathbb{E}X_s < 1$ and $\rho_f = \lambda_f\mathbb{E}X_f < 1$.

Optimal slow server selection probability

Minimizing mean task completion time

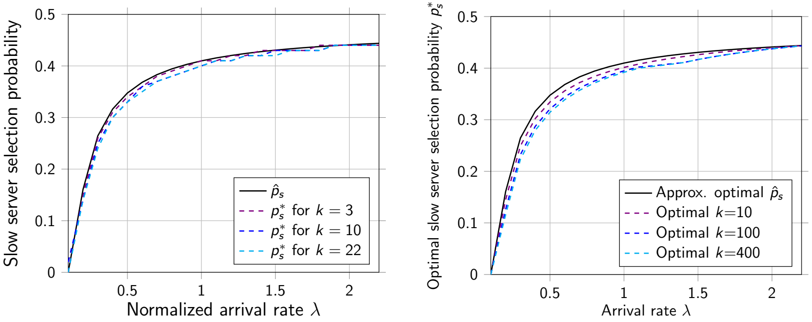
- Find optimal probability p_s^* that minimizes $\mathbb{E}T$ difficult to compute analytically
- Find probability $\hat{p}_s$ that minimizes both the upper and the lower bound
- Approximate p_s^* by $\hat{p}_s$

Exponentially distributed sub-task completion times

$$\hat{p}_s = \begin{cases} 0, & \lambda \leqslant \tau_1, \\ \frac{1-\frac{\tau_1}{\lambda}}{1+\frac{\bar{f}_s}{f_s}\sqrt{\frac{\mu_f}{\mu_s}}}, & \tau_1 \leqslant \lambda < \mu_s f_s + \mu_f \bar{f}_s. \end{cases}$$

where $\tau_1 \triangleq \bar{f}_s(\mu_f - \sqrt{\mu_s \mu_f})$

Slow server selection probabilities

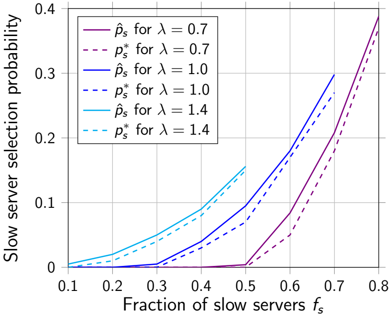


(a) $n = 10^2, k \in \{3, 10, 22\}$.

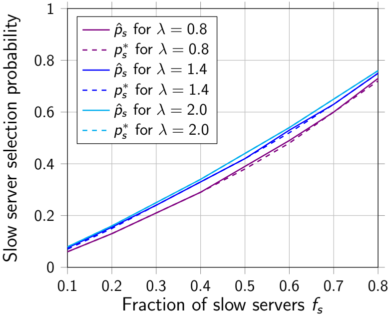
(b) $n = 10^3, k \in \{10, 100, 400\}$.

Figure: System with $(\mu_s, \mu_f) = (2, 2.5)$.

Variation with the fraction of slow servers f_s



(a) $(\mu_s, \mu_f) = (0.5, 2.5)$.



(b) $(\mu_s, \mu_f) = (2.0, 2.5)$.

Figure: System with $n = 10^3$, $k = 10$

Deterministic Scheduling

Deterministic selection

Select k_s slow and $k - k_s$ fast servers

Theorem

The optimal selection probability of slow servers, converges to

$$\lim_{k\to\infty} p_s^* = \frac{k_s^*}{k},$$

where k_s^* is the optimal deterministic selection of slow servers. The optimal probability of choosing ℓ servers converges to

$$\lim_{k\to\infty} q(\ell) = \mathbb{1}_{\{\ell = kp_s^*\}}.$$

Mean number of slow servers

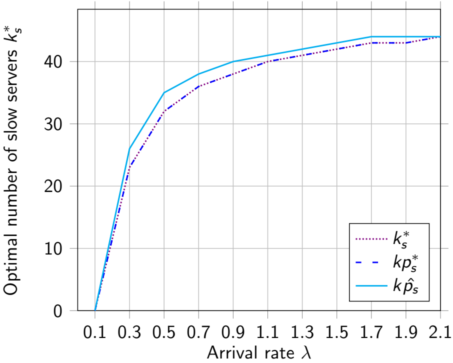
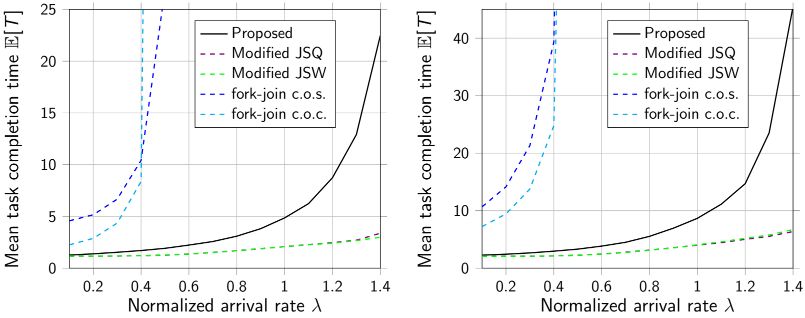


Figure: System with $n = 10^3$, $k = 100$, $f_s = 0.5$, $(\mu_s, \mu_f) = (2, 2.5)$.

Comparison with other Load Balancing Policies



(a) $n = 10^2,\ k = 10$

(b) $n = 10^3,\ k = 10^2$

Figure: System with $f_s = 0.5$, $(\mu_s, \mu_f) = (0.5, 2.5)$

Conclusion

- We show that the joint distribution of the stationary workload across k queues becomes asymptotically independent as the number of servers, n, grows and $k = o(n^{\frac{1}{4}})$
- We derive an upper and lower bound on the limiting mean response time and identify the selection probability, p_s, that minimizes this bound
- Numerical experiments confirm that the selected probability provides near-optimal performance

Acknowledgements

References

- R. Jinan, G. Gautam, V. Aggarwal, P. Parag. Asymptotic analysis of probabilistic scheduling for erasure-coded heterogeneous systems. ACM SIGMETRICS 2023.

Centre for Networked Intelligence (CNI)

Activities

Infrastructure

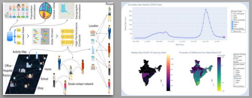
- Core Research
- Applied R &D
- Teaching & training
- Outreach

Outcomes

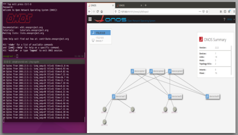
Core Research	Publications	Peer-reviewed papers in top-tier conferences/journals, pre-prints
Core Research, Applied R &D	Policy inputs	KSDMA, Government of Karnataka, DST, Ministry of Housing and Urban Affairs, Government of India, Smart City, ITU
Applied R &D	PoC solutions/ demos	
Teaching & training	Scholarships, internships	PhD & MTech fellowships, Postdoc, internships
Teaching & training	Courses, labs, FDP	Online courses, FDP, cloud-based lab (IISc/public use)
Outreach	Knowledge dissemination	Workshops, Summer School
Outreach	Community-building Awareness	Conferences, seminars, hackathons DIGITAL DEFENDERS CYBER SECURITY MASTERCLASS SERIES & CTF

CENTRE FOR NETWORKED INTELLIGENCE
Indian Institute of Science

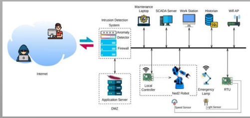
CNI R & D Projects



Pandemic Response

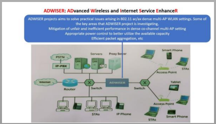


Data Center & Cloud

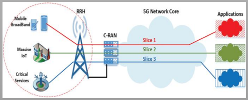


Industrial Control Systems Security

Tactile CyberPhysical Systems Testbed



ADWISER: WiFi Optimisation



Cellular Network Slicing

CENTRE FOR
NETWORKED INTELLIGENCE
Indian Institute of Science

Webpage

For more details: `https://cni.iisc.ac.in`

GATE

GATE opens avenues for

- Admission to Masters and PhD programs in prestigious institutes across the country and in some places abroad.
- Financial support for MoE and AICTE-approved PG programs in India
- Job opportunities in Public Sector Units and government research organizations