Tackling Heterogeneous Traffic in Multi-access Systems via Erasure Coded Servers

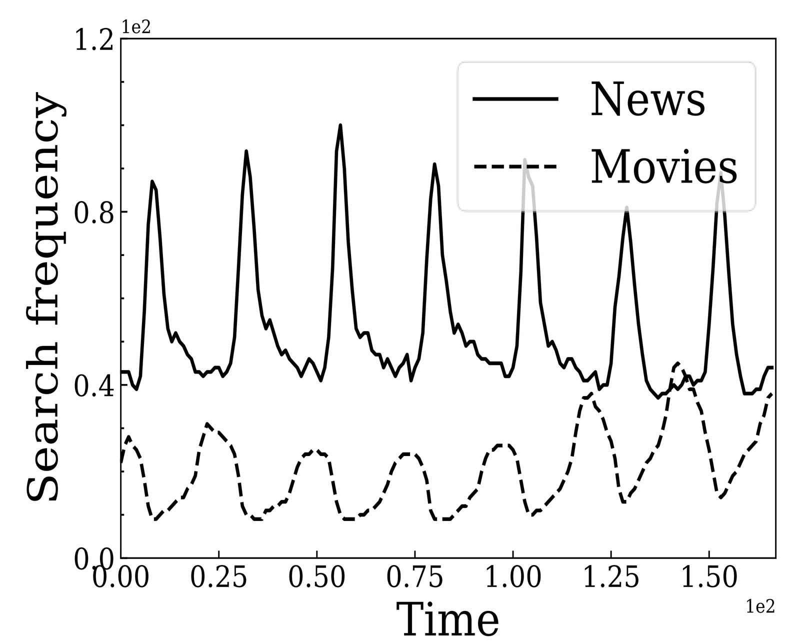
Tuhinangshu Choudhury, Weina Wang, Gauri Joshi

Carnegie Mellon University

SIGMETRICS acm

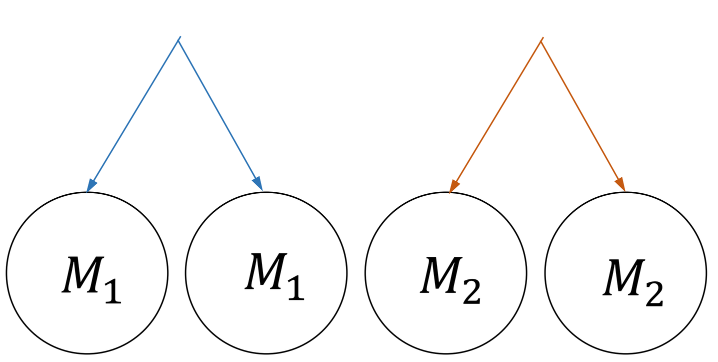
Motivation

- Servers in cloud systems allocated based on traffic.
- Traffic can vary significantly.
- Over-provisioning of servers to maintain QoS.
- Under-utilization of resources during low traffic.



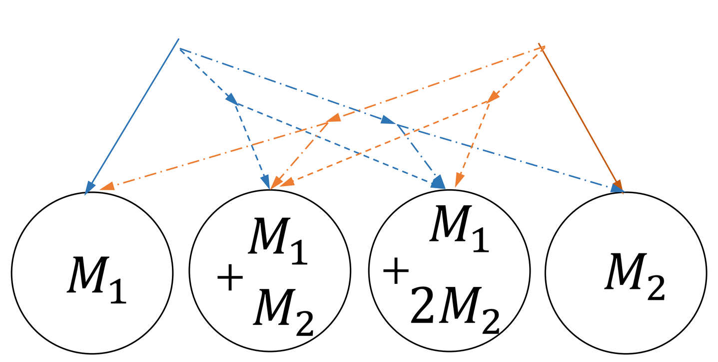
A Simple Example

- Cloud system storing two matrices M_1 and M_2.
- 4 unit rate servers processing download requests.



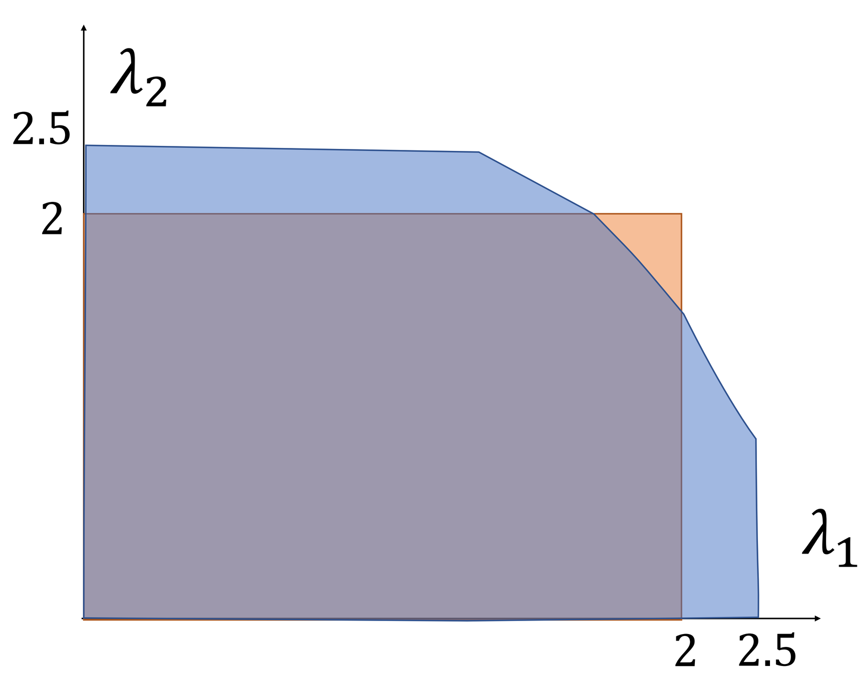
Traditional System

Contain only systematic servers



Coded System

Contain systematic and coded servers

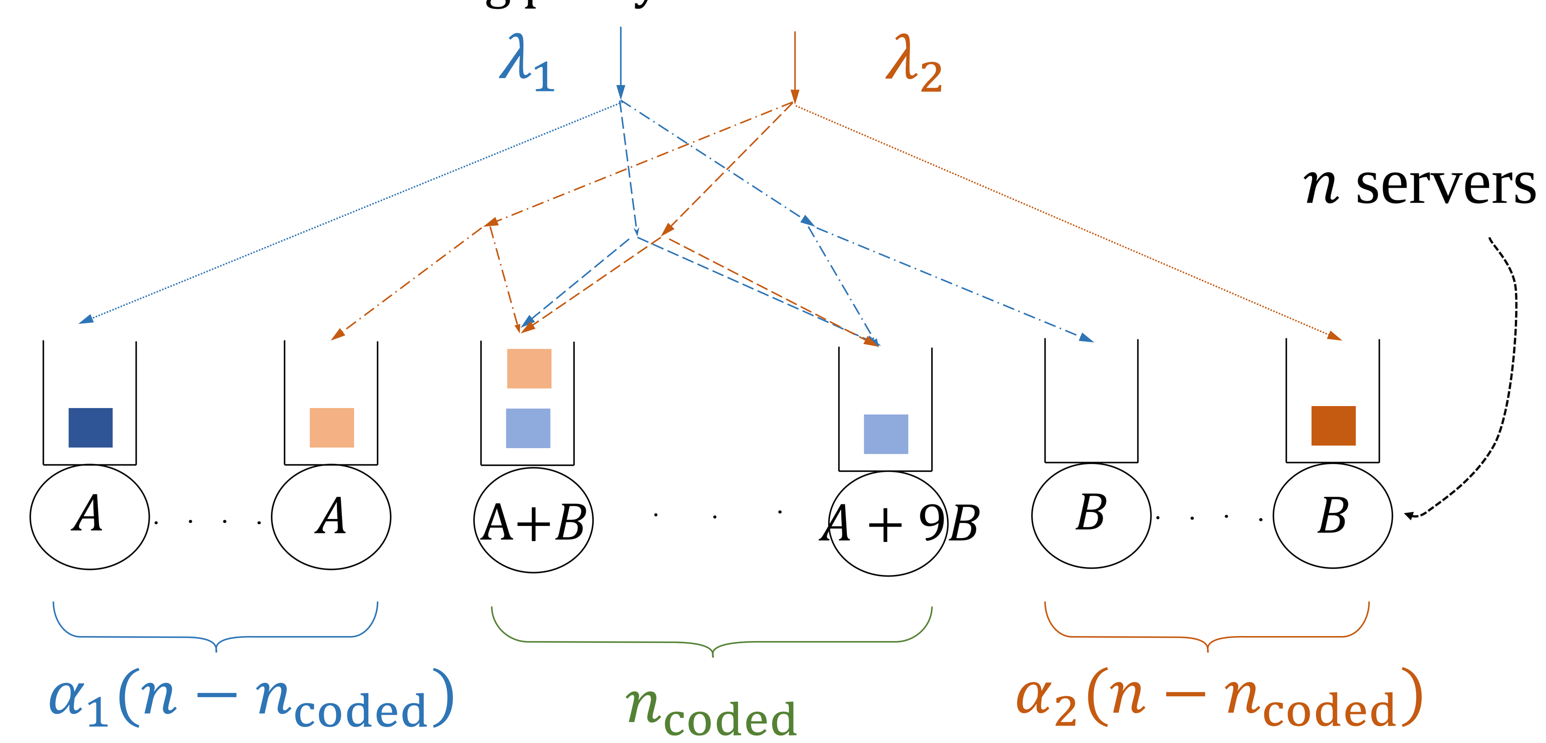


Key Difference: Systematic servers can serve only one job type, while coded servers can serve multiple job types when combined with systematic servers.

- Traditional system offers one way to download a file, while the coded system provides three.
- Increases maximum serviceable rate from 2 to 2.5.
- Coded servers improve flexibility, thus allowing usage of under-utilized systematic servers.

System Model

- k job types with arrival according to $PP(\lambda_i)$ for the i^{th} type.
- n servers with $Exp(1)$ service .
- α_i fraction of systematic servers is allocated for the i^{th} type.
- Replace n_{coded} systematic servers with erasure-coded servers.
- Probabilistic routing policy to choose a combination of servers.



Service Capacity Region

Compared to previous work, we extended the service capacity region characterization for any general k and reduced the stability criteria from a combinatorial equation to a linear equation.

Theorem: Service Capacity Region Characterization

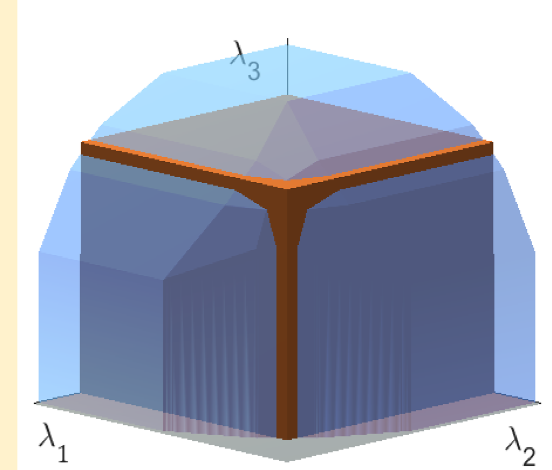
$r_i \triangleq \alpha_i(n - n_{\text{coded}}) - \lambda_i$

$r_i^+ = \max(r_i, 0)\,, r_i^- = -\min(r_i, 0)$

W.l.o.g $r_1 \le r_2 \le \cdots \le r_k$. Then $(\lambda_1, \ldots, \lambda_k) \in \Lambda_{\text{coded}}$ iff

$$\underbrace{\min_{k_0 \in \{1,\ldots,k\}} \left(\frac{n_{\text{coded}} + \sum_{i=1}^{k_0} r_i^+ \}}{k_0} \right)}_{\text{Extra Service Capacity}} \ge \underbrace{\sum_{i}^{k} r_i^-}_{\text{Extra Service Requirement}}$$

For k = 2, $\Theta(nn_{\text{coded}})$ increase in volume of SCR



Key Technique: A water-filling algorithm that first utilizes the systematic servers to the fullest before using the coded servers.

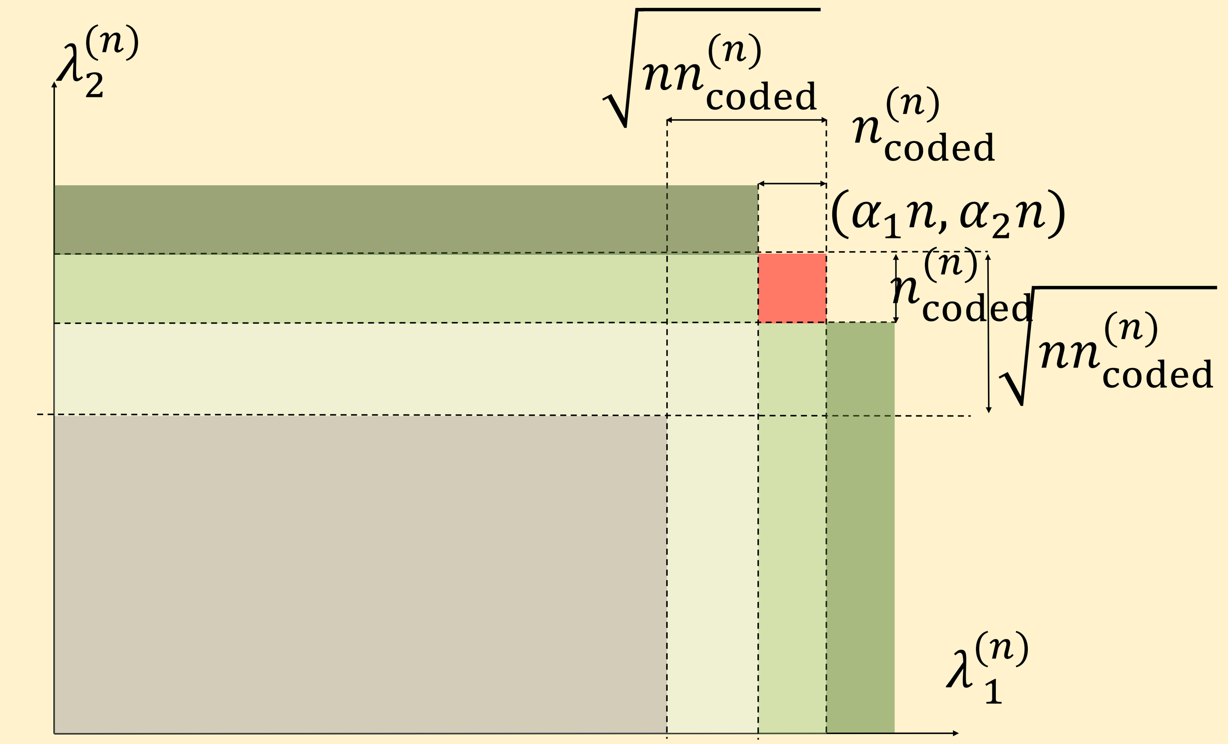
Summary

- Few erasure-coded servers significantly improve the service capacity region.
- In terms of mean response time, the coded system is significantly better than or at least as good as the uncoded system in most regimes.
- Coded system is more resilient to variation in traffic.

Response Time

We characterize the difference in response time for the coded system and the uncoded system for different traffic regimes.

Theorem: Response Time Characterization



- Light: $\left|\mathbb{E}\left[T_{\text{coded}}^{(n)}\right] - \mathbb{E}\left[T_{\text{uncoded}}^{(n)}\right]\right| = o(1)$
- Inner-heavy: $\mathbb{E}\left[T_{\text{coded}}^{(n)}\right] \le \mathbb{E}\left[T_{\text{uncoded}}^{(n)}\right] - \omega(1)$
- Outer-heavy: $\mathbb{E}\left[T_{\text{coded}}^{(n)}\right] = o\left(\mathbb{E}\left[T_{\text{uncoded}}^{(n)}\right]\right)$
- Uncoded-unstable: $\mathbb{E}\left[T_{\text{coded}}^{(n)}\right] < \infty, \mathbb{E}\left[T_{\text{uncoded}}^{(n)}\right] = \infty$
- Coded-unstable: $\mathbb{E}\left[T_{\text{coded}}^{(n)}\right] = \infty, \mathbb{E}\left[T_{\text{uncoded}}^{(n)}\right] < \infty$

Key Technique: Designing an optimal routing policy that reduces the load from the heavier loaded systematic servers while maintaining the system's stability.

Regimes for $k = 2$

- Light: $\beta_i^{(n)} = \omega\left(\sqrt{nn_{\text{coded}}^{(n)}}\right)$ for all i
- Inner-heavy: $\beta_1^{(n)} \in \left(\Omega\left(n_{\text{coded}}^{(n)}\right), o\left(\sqrt{nn_{\text{coded}}^{(n)}}\right)\right), \beta_2^{(n)} = \omega\left(\beta_1^{(n)}\right)$
- Outer-heavy: $\beta_1^{(n)} = o\left(n_{\text{coded}}^{(n)}\right), \beta_2^{(n)} = \omega\left(n_{\text{coded}}^{(n)}\right)$

For regime definitions in the case of any arbitrary k, refer to the paper.

Experiments

Two job types with arrival rate in form of a pulse wave