Capacity of Distributed SDC Resources

Pengchao Han (Imperial), Shiqiang Wang (IBM US), Kin K. Leung (Imperial), Kevin Chan (ARL), Don Towsley (UMass)

Distributed SDC Resources

How to characterize capacity to enable efficient use of distributed SDC resources belonging to different coalition members?

Resource Capacity

Consider a system (SDC slice) with multiple and distributed resources running multiple tasks

- K types of resources
- N servers: normalized capacity $c_{n,k} = 1, \forall n,k$
- M tasks: random resource requirements $X_{m,k}$
  - Requirements of different resources for a given task are independent with each other
  - Occupancy of resource type k on server n, $\rho_{n,k} = \sum_{m=1}^{M} X_{m,k} I_{m,n}$ where $I_{m,n}$=1 if task m is assigned to server n, and 0 otherwise

Definition of resource capacity: The maximum number of tasks M that the system can serve simultaneously, so that the overload probability is no more than a small value ε:

$$\Pr\left\{\bigcup_{k=1}^{K}\bigcup_{n=1}^{N}\left(\rho_{n,k} \geq 1\right)\right\} \leq \varepsilon$$

✓ **Capacity of distributed resources can be used to assist scheduling or admission decisions of distributed analytics**

Main Analytical Results

Let $\mu^{\max}$ and $\left(\sigma^{\max}\right)^2$ denote the maximum mean and variance among all required resources for all tasks. We consider 3 assignment policies for scenarios with different resource monitoring overheads.

Random Assignment (RAND)

Assign M tasks to N servers randomly:

$$M \leq \frac{N\left(\log\frac{\varepsilon}{NK} + \theta\right)}{\exp\left(\mu^{\max}\theta + \left(\sigma^{\max}\theta\right)^2/2\right) - 1}$$

Power of d Choices Assignment (PODC)

For each task, randomly choose d servers and assign it to the least loaded one among them:

$$M \leq \frac{N\log\frac{\varepsilon}{NK} + N\theta}{\mu^{\max}\theta + \left(\sigma^{\max}\theta\right)^2/2} - \frac{N\log\log N}{\log d}$$

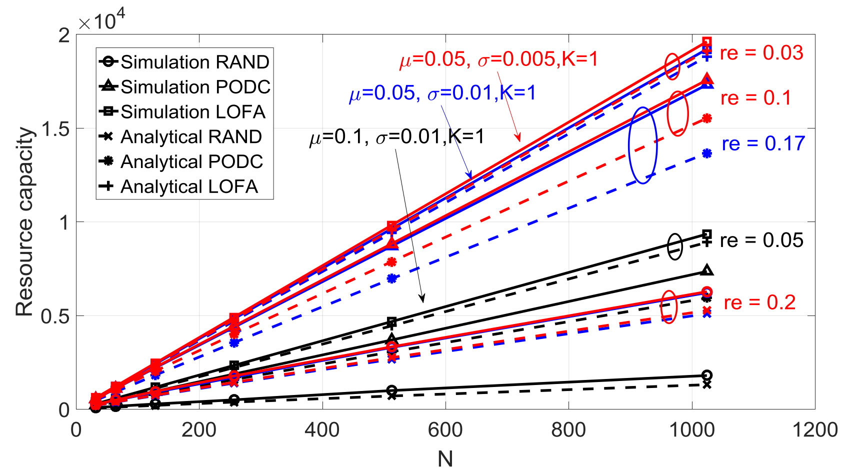
Least Occupancy First Assignment (LOFA)

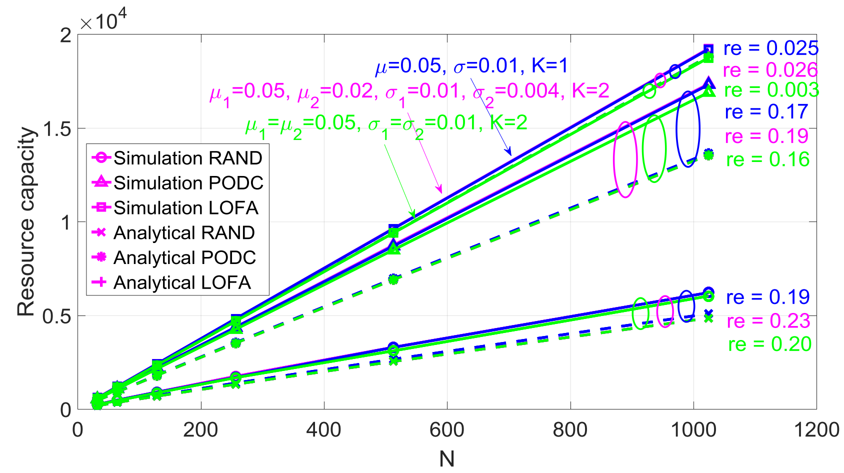
Assign each task to the server with the least occupancy (loading):

$$M \leq \frac{\log\varepsilon + \left(N - N\mu^{\max}\right)\theta - \left(N\sigma^{\max}\theta\right)^2/2}{\mu^{\max}\theta + \left(\sigma^{\max}\theta\right)^2/2} + 1$$

- *Gaussian distributed resource requirements are assumed in the above expressions*
- *We can find the optimal θ^* to obtain the better upper-bound for M*
- *The bounds above are sufficient conditions for guaranteeing ε*
- *The analysis also allows other distributions*

Numerical Results





Tightness of resource capacities with different μ, σ, K ($\varepsilon = 0.01, d = 2$), "re" stands for mean relative error