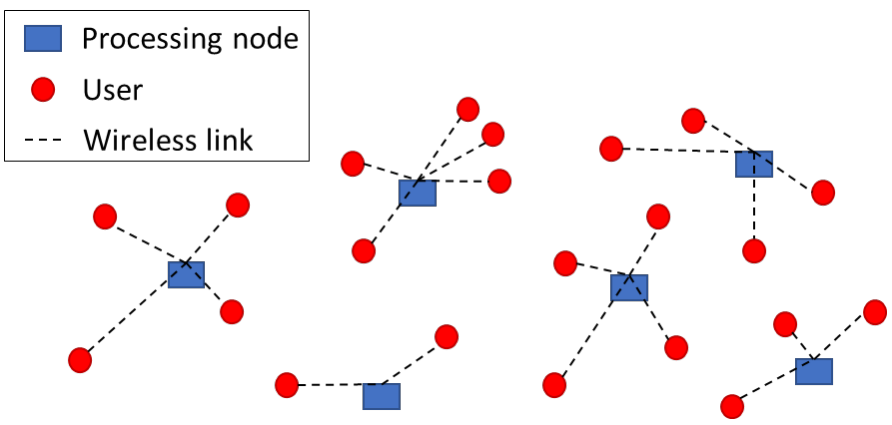
Analytics Capacity: Theoretical Formulation and Scaling Laws

Shiqiang Wang (IBM US), **Ting He** (PSU), **Prithwish Basu** (BBN), **Theodoros Salonidis** (IBM US), **Kevin Chan** (ARL)

Distributed Analytics



System with multiple processing nodes running analytics applications

- N processing nodes
- M users (equivalent to active application requests)
- Each user receives service from its closest processing node
  - We also refer to a processing node as a **cell**
  - The probability that a user is in cell i is $\frac{1}{N}$ and the cell associations of all users are *i.i.d.*
- L different applications
  - Each user chooses an application with probability $\frac{1}{L}$
  - The application choice of different users are *i.i.d.*
- Service capacity: Each cell can host at most R different applications and serve at most K users
- Proactive service replication: A user's application must exist in the user's cell and γ of its neighboring cells
  - To allow smooth transition when the user moves to another cell of high probability

Analytics Capacity: Definition

➢ The maximum number of users M that the system can support, so that the probability of *at least one* cell exceeding its service capacity (constrained by R and K) approaches zero as $N \to \infty$.

Analytics Capacity Result

Lower Bound (Sufficient Condition)

For any $\alpha > 1$, if

$$M \le min\left\{KN^{1-\frac{\alpha}{K}}; \frac{RN^{1-\frac{\alpha}{R}}}{1+\gamma} \cdot \frac{L}{L-R}\right\}$$

then the probability that the system exceeds its service capacity in at least one cell is at most $\epsilon = \frac{e^K + \left(\frac{Le}{L-R}\right)^R}{N^{\alpha-1}}$.

As $N \to \infty$, $\epsilon \to 0$, the *achievable lower bound of analytics capacity* is

$$\Omega\left(N^{1-\frac{\alpha}{min\{K,R\}}}\right)$$

Upper Bound (Necessary Condition)

If the probability that the system exceeds its capacity in at least one cell is at most ϵ, when $M > K$, $L > R$, and the value of ϵ is chosen small enough such that $1 - \epsilon N > 0$ and $\epsilon < 1$, then we have

$$\mathrm{M} \le min\left\{\frac{(1-\epsilon)NK}{1-\epsilon N}; \frac{log\left((1-\epsilon)\left(1-\frac{R}{L}\right)\right)}{\log\left(1-\frac{(1+\gamma)}{NL}\right)}\right\}$$

When $\epsilon = 0$, as $N \to \infty$, the *analytics capacity upper bound* is

$$O(N)$$

❖ For sufficiently large K and R, the asymptotic gap between the lower and upper bounds is small

❖ Insight: When expanding system capacity, one should simultaneously increase the number of processing nodes and the capability of each node