Reconfigurable Intelligent Surface for Green Edge Inference in Machine Learning

Sheng Hua, Yuanming Shi

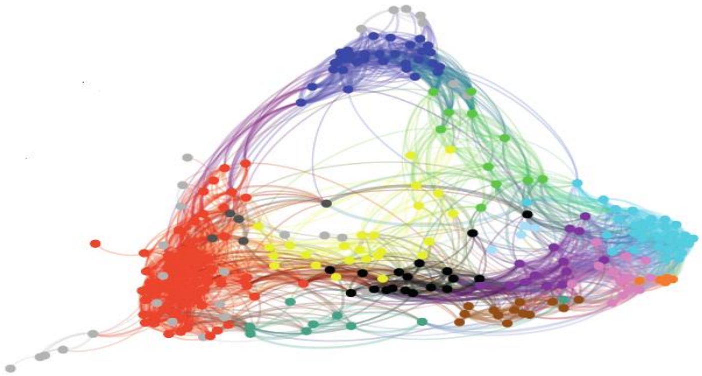
ShanghaiTech University

上海科技大学
ShanghaiTech University

Outline

- **Motivations**
  - Storage, latency, power
- **Two vignettes:**
  - **Energy-efficient edge cooperative inference**
    - Why inference at network edge?
    - Edge inference via wireless cooperative transmission
  - **Reconfigurable intelligent surface empowered edge inference**
    - Why reconfigurable intelligent surface?
    - Joint phase shifts and beamforming vectors design

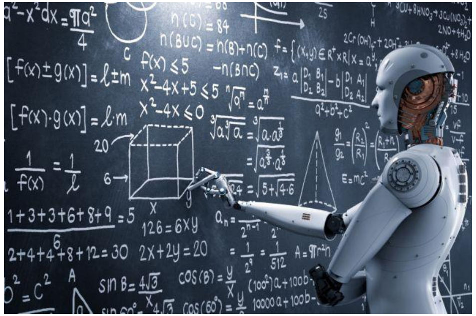
Vignettes A: Energy-efficient edge cooperative inference



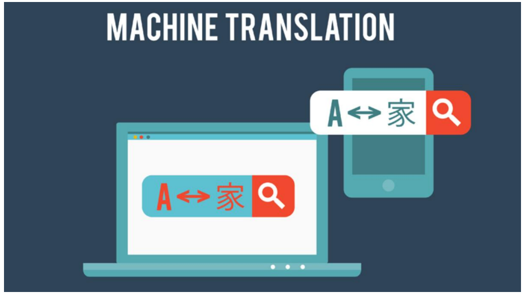
Why edge inference?

AI is changing our lives

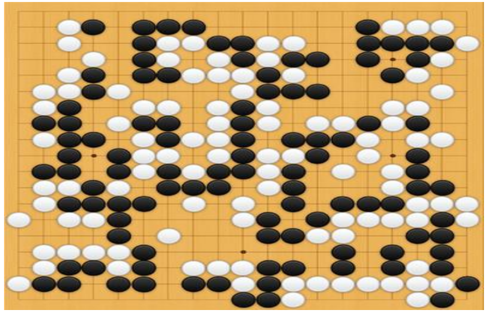
self-driving car



smart robots



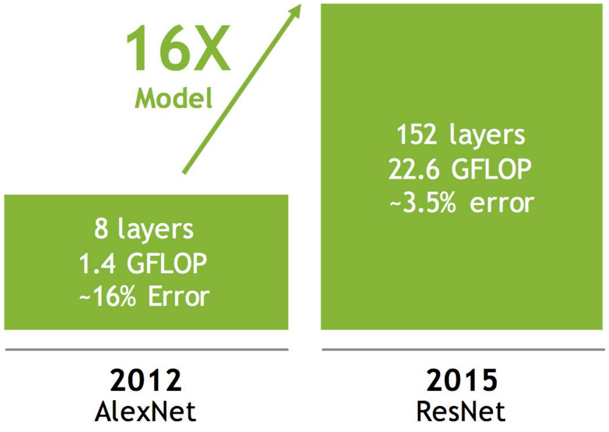
machine translation



AlphaGo

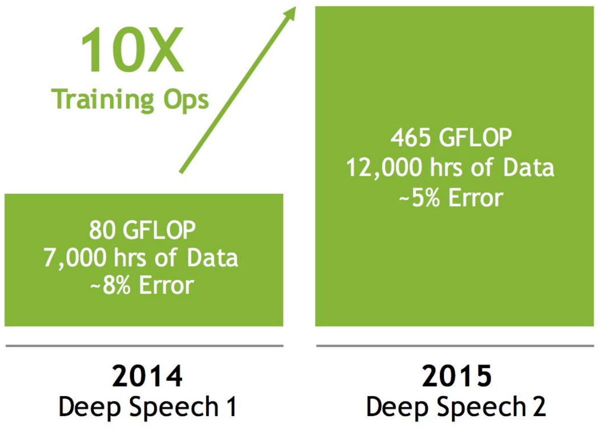
Models are getting larger

image recognition



Microsoft

speech recognition



Bai du 百度

Fig. credit: Dally

The first challenge: model size

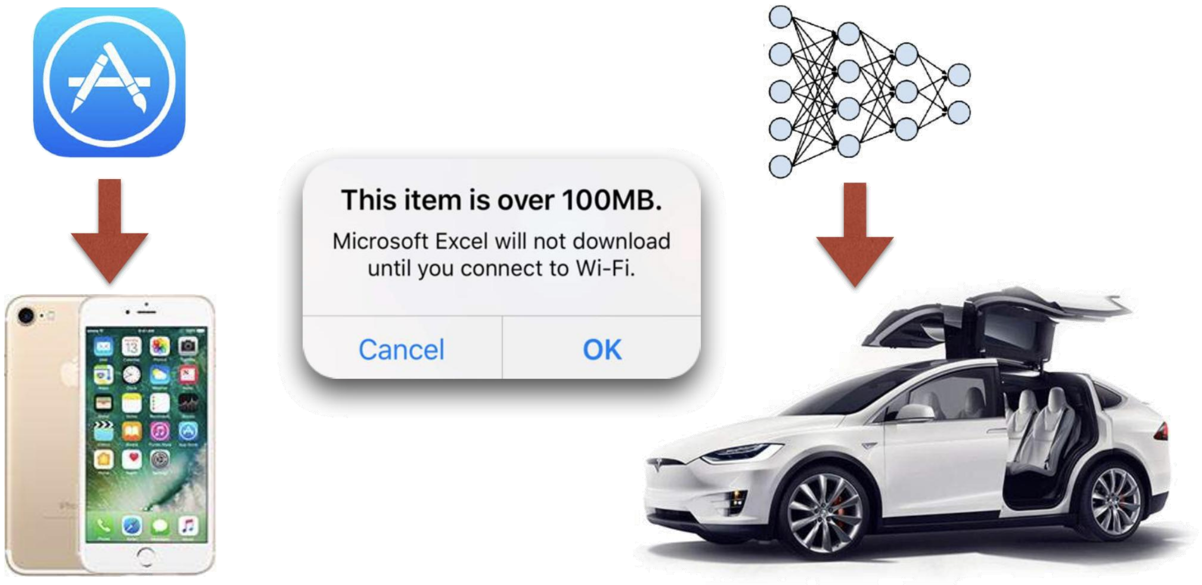


Fig. credit: Han

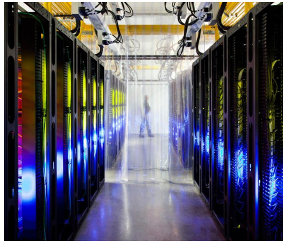
difficult to distribute large models through over-the-air update

The second challenge: energy

AlphaGo: 1920 CPUs and 280 GPUs,

$3000 electric bill per game

on mobile: drains battery



larger model-more memory reference-more energy

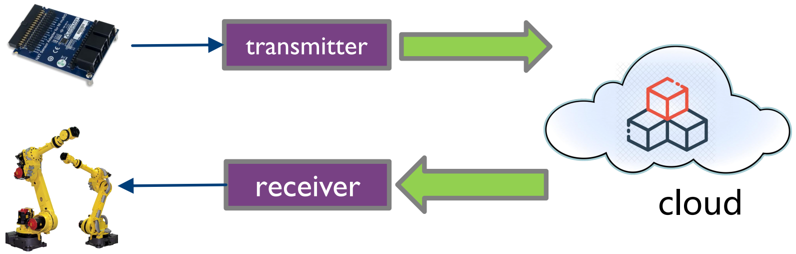
The third challenge: speed

	Error rate	Training time
ResNet18:	10.76%	2.5 days
ResNet50:	7.02%	5 days
ResNet101:	6.21%	1 week
ResNet152:	6.16%	1.5 weeks

long training time limits ML researcher's productivity

communication

sensor



actuator

latency

processing at "Edge" instead of the "Cloud"

How to make deep learning more energy-efficient?

low power

Edge inference for deep neural networks

- **Goal:** energy-efficient edge processing framework to perform deep learning inference tasks at the edge computing nodes

any task can be performed at multiple BSs

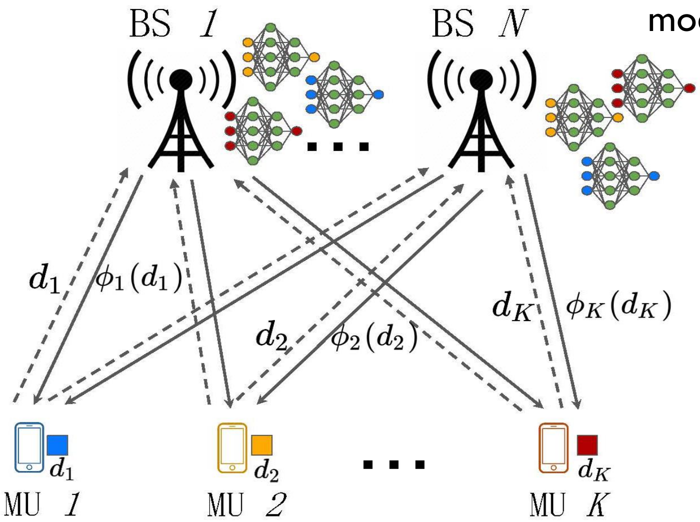
uplink

downlink

models

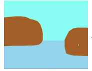
models

pre-downloaded



which BSs shall compute for me?

output



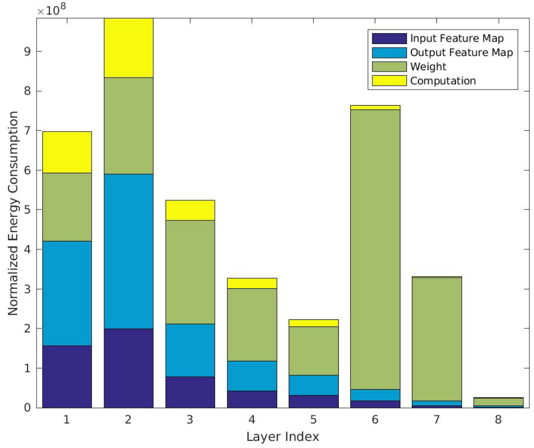
input

example:

Nvidia's GauGAN

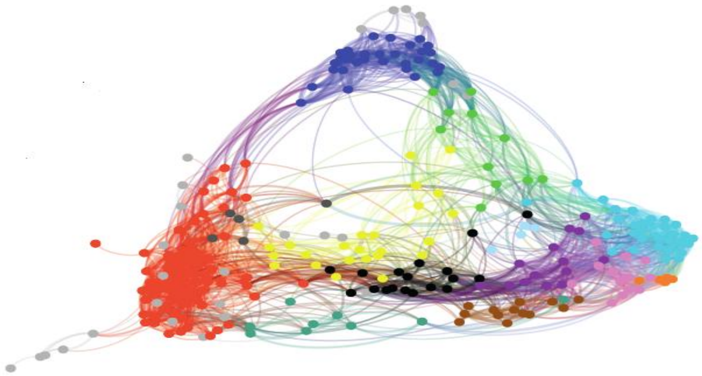
Computation power consumption

- **Goal:** estimate the power consumption for deep model inference
- Example: power consumption estimation for AlexNet [Sze' CVPR 17]



- Cooperative inference tasks at multiple BSs:
  - *Computation replication:* high computation power
  - *Cooperative transmission:* low transmission power
- **Solution:**
  - minimize the sum of computation and transmission power consumption

Vignettes B: Reconfigurable intelligent surface empowered edge inference



Smart radio environments

- Current wireless networks: no control of radio waves
  - Perceive the environment as an "unintentional adversary" to communication
  - Optimize only the end-points of the communication network
  - No control of the environment, which is viewed as a passive spectator
- Smart radio environments: reconfigure the wireless propagations

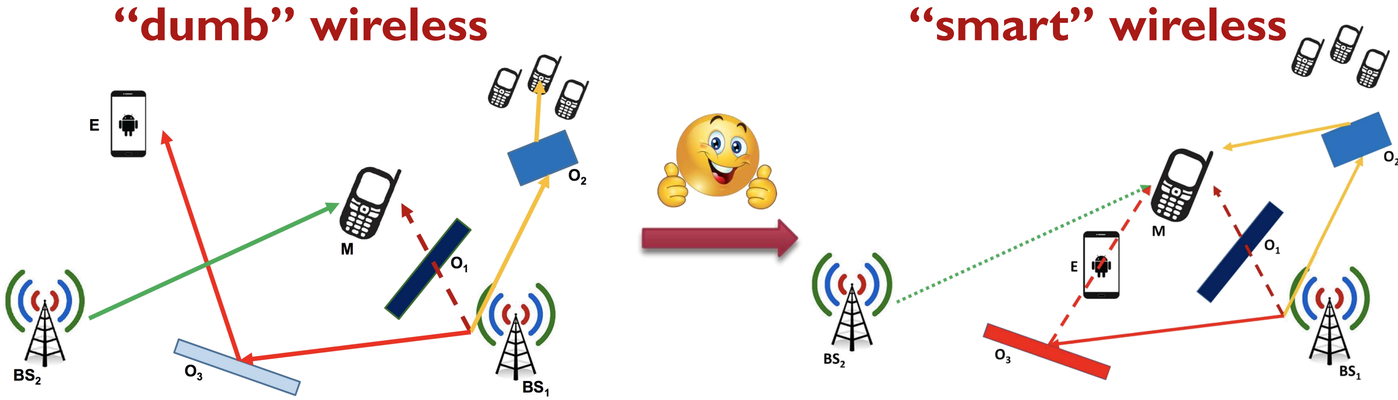


Fig. credit: Renzo

Reconfigurable intelligent surface

- **Working principle of reconfigurable intelligent surface (RIS):** different elements of an RIS can reflect the incident signal by controlling its amplitude and/or phase for directional signal enhancement or nulling

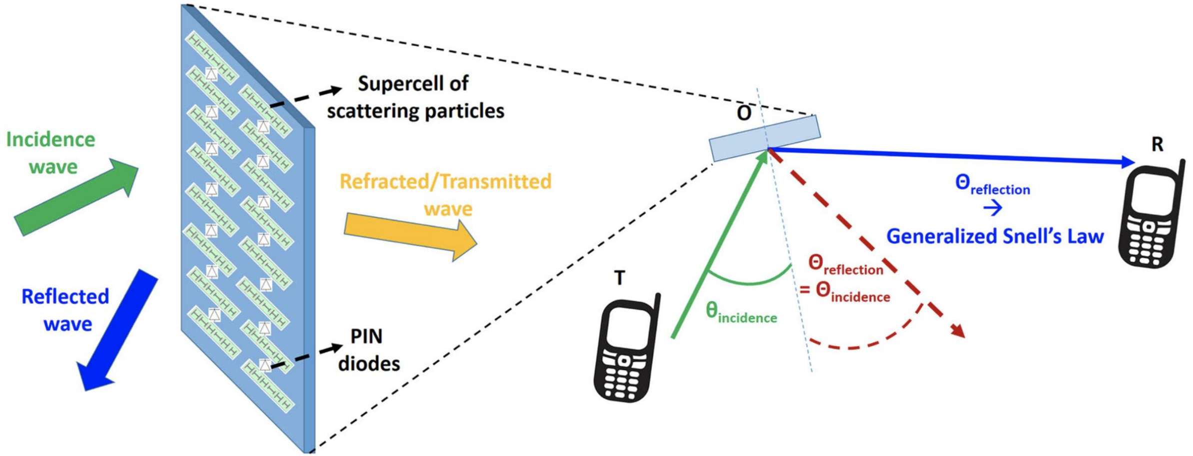
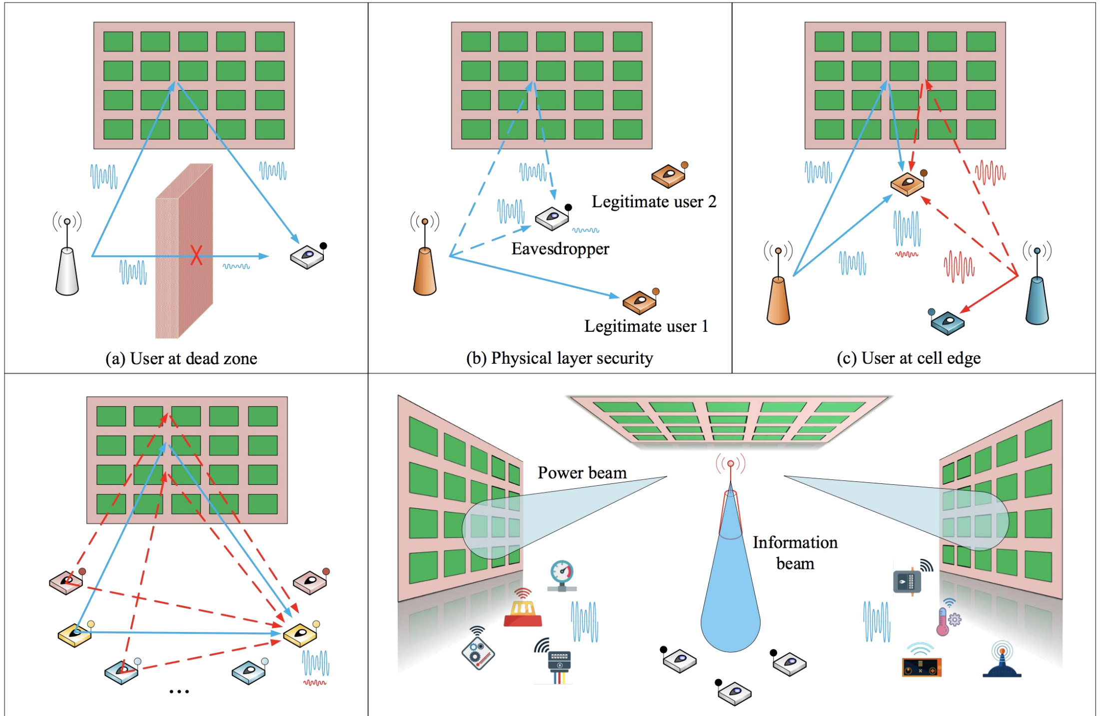


Fig. credit: Renzo

1. no any active transmit module
2. operate in full-duplex mode

improve spectral and energy efficiency

Reconfigurable intelligent surface meet wireless networks



(a) User at dead zone (b) Physical layer security (c) User at cell edge (d) Massive D2D communications (e) Wireless information and power transfer in an IoT network

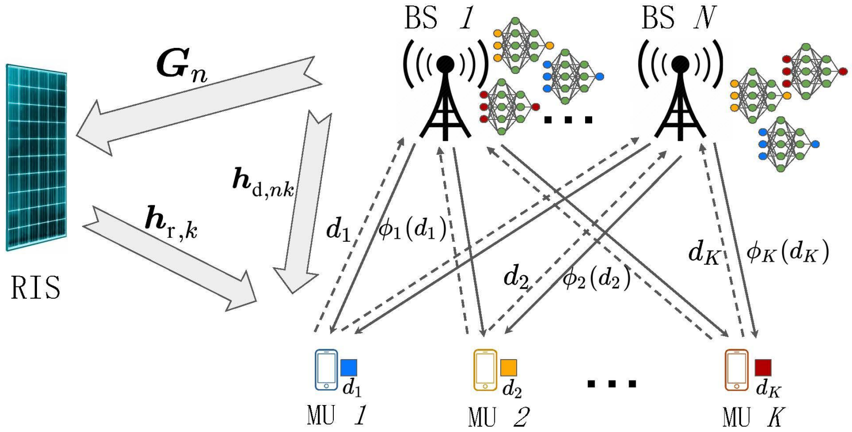
reconfigurable intelligent surface meets wireless network:

- edge inference
- over-the-air computation
- massive MIMO
- wireless power transfer
- D2D communications
- NOMA
- mmWave
- …

Fig. credit: Wu

RIS empowered edge inference

- Reconfigurable Intelligent surface:
  - overcoming unfavorable signal propagation conditions
  - improving energy efficiency
  - tuning phase shifts with M passive elements



$$\boldsymbol{\Theta} = \mathrm{diag}(\beta\theta_1, \cdots, \beta\theta_M)$$
$$\text{with } \theta_m = e^{j\varphi_m}, \varphi_m \in [0, 2\pi)$$

w.l.o.g. assuming $\beta = 1$

RIS aided edge inference system:
build controllable wireless environments to decrease transmit signal power

Signal model

- **Proposal:** MU k's task performed at multiple BSs $\mathcal{A}_n \subseteq \mathcal{K}$
  - transmitted signal at BS n: $\boldsymbol{x}_n = \sum_{k \in \mathcal{A}_n} \boldsymbol{v}_{nk} s_k$
  - beamforming vector for $\phi_k(d_k)$ at BS n: $\boldsymbol{v}_{nk}$
  - signal received by MU $k \in \mathcal{K}$: $y_k = \sum_{n \in \mathcal{N}} \boldsymbol{g}_{nk}^{\mathrm{H}} \boldsymbol{x}_n + z_k$
  - equivalent channel response from BS n to MU k :

$$\boldsymbol{g}_{nk} = \underbrace{\boldsymbol{h}_{\mathrm{d},nk}}_{\text{direct link}} + \underbrace{\boldsymbol{G}_n^{\mathrm{H}} \boldsymbol{\Theta}^{\mathrm{H}} \boldsymbol{h}_{\mathrm{r},k}}_{\text{reflected link}}$$

  - the SINR for MU $k \in \mathcal{K}$:

$$\mathrm{SINR}_k(\mathcal{A}) = \frac{\left|\sum_{n \in \mathcal{N}} \mathbf{1}_{\{k \in \mathcal{A}_n\}} \boldsymbol{g}_{nk}^{\mathrm{H}} \boldsymbol{v}_{nk}\right|^2}{\sum_{l \neq k} \left|\sum_{n \in \mathcal{N}} \mathbf{1}_{\{l \in \mathcal{A}_n\}} \boldsymbol{g}_{nk}^{\mathrm{H}} \boldsymbol{v}_{nl}\right|^2 + \sigma_k^2}$$

Energy-efficient edge inference

- **Goal:** minimize total power consumption under QoS constraints

$$
\begin{aligned}
\mathscr{P}_{\text{original}} : \underset{\mathcal{A},\{\boldsymbol{v}_{nk}\},\boldsymbol{\Theta}}{\text{minimize}} \quad & \sum_{n\in\mathcal{N}} \frac{1}{\eta_n} \sum_{k\in\mathcal{A}_n} \|v_{nk}\|_2^2 + \sum_{n\in\mathcal{N}} \sum_{k\in\mathcal{A}_n} P_{nk}^{\mathrm{c}} \\
\text{subject to} \quad & \mathrm{SINR}_k(\mathcal{A}) \geq \gamma_k, \quad \forall\, k \in \mathcal{K}, \\
& \sum_{k\in\mathcal{A}_n} \|\boldsymbol{v}_{nk}\|_2^2 \leq P_{n,\max}, \quad \forall\, n \in \mathcal{N}, \\
& |\theta_m| = 1, \quad \forall\, m \in \mathcal{M},
\end{aligned}
$$

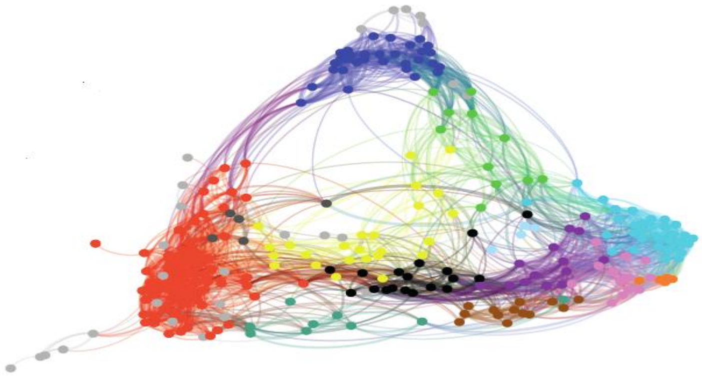
sum of communication and computation power consumption

(maximum transmit power)

phase shifts design

- **Challenges:**
  - 1. mixed combinatorial optimization problem because of combinatorial variable $\mathcal{A} = (\mathcal{A}_1, \ldots, \mathcal{A}_N)$
  - 2. coupled optimization variables in SINR constraints
  - 3. nonconvex unit-modulus constraints induced by the RIS

Group Sparsity Inducing and An Alternating Framework



Group sparse beamforming for power minimization

- **Proposal:** group sparse beamforming approach to get rid of the combinatorial variable $\mathcal{A}$
- **Key observation:** $k \notin \mathcal{A}_n \Leftrightarrow \boldsymbol{v}_{nk} = \mathbf{0}$

$$\sum_{n\in\mathcal{N}} \sum_{k\in\mathcal{A}_n} P_{nk}^{\mathrm{c}} \Rightarrow \sum_{n\in\mathcal{N}} \sum_{k\in\mathcal{K}} \mathbf{1}_{\{\boldsymbol{v}_{nk}=\mathbf{0}\}} P_{nk}^{\mathrm{c}}$$

$$\mathrm{SINR}_k(\mathcal{A}) = \frac{\left|\sum_{n\in\mathcal{N}} \mathbf{1}_{\{k\in\mathcal{A}_n\}} \boldsymbol{g}_{nk}^{\mathrm{H}} \boldsymbol{v}_{nk}\right|^2}{\sum_{l\neq k} \left|\sum_{n\in\mathcal{N}} \mathbf{1}_{\{l\in\mathcal{A}_n\}} \boldsymbol{g}_{nk}^{\mathrm{H}} \boldsymbol{v}_{nl}\right|^2 + \sigma_k^2}$$

$$\Rightarrow \mathrm{SINR}_k = \frac{\left|\sum_{n\in\mathcal{N}} \boldsymbol{g}_{nk}^{\mathrm{H}} \boldsymbol{v}_{nk}\right|^2}{\sum_{l\neq k} \left|\sum_{n\in\mathcal{N}} \boldsymbol{g}_{nk}^{\mathrm{H}} \boldsymbol{v}_{nl}\right|^2 + \sigma_k^2}, \text{ where } \boldsymbol{v}_{nk} = \mathbf{0} \text{ if } k \notin \mathcal{A}_n$$

Group sparse beamforming for power minimization

- **Proposal:** exploit group sparsity structure beamforming to get rid of the combinatorial variable $\mathcal{A}$

$$
\begin{aligned}
\mathscr{P}_{\text{original}}: \underset{\mathcal{A},\{\boldsymbol{v}_{nk}\},\boldsymbol{\Theta}}{\text{minimize}} \quad & \sum_{n\in\mathcal{N}} \frac{1}{\eta_n} \sum_{k\in\mathcal{A}_n} \|v_{nk}\|_2^2 + \sum_{n\in\mathcal{N}} \sum_{k\in\mathcal{A}_n} P_{nk}^{\mathrm{c}} \\
\text{subject to} \quad & \mathrm{SINR}_k(\mathcal{A}) \geq \gamma_k, \quad \forall\, k \in \mathcal{K}, \\
& \sum_{k\in\mathcal{A}_n} \|\boldsymbol{v}_{nk}\|_2^2 \leq P_{n,\max}, \ \forall\, n \in \mathcal{N}, \\
& |\theta_m| = 1, \quad \forall\, m \in \mathcal{M},
\end{aligned}
$$

$$k \notin \mathcal{A}_n \Leftrightarrow \boldsymbol{v}_{nk}^{\mathrm{DL}} = \mathbf{0}$$

$$
\begin{aligned}
\underset{\{\boldsymbol{v}_{nk}\},\boldsymbol{\Theta}}{\text{minimize}} \quad & \sum_{n\in\mathcal{N}} \sum_{k\in\mathcal{K}} \frac{1}{\eta_n} \|\boldsymbol{v}_{nk}\|_2^2 + \sum_{n\in\mathcal{N}} \sum_{k\in\mathcal{K}} \mathbf{1}_{\{\boldsymbol{v}_{nk}=\mathbf{0}\}} P_{nk}^{\mathrm{c}} \\
\text{subject to} \quad & \mathrm{SINR}_k \geq \gamma_k, \quad \forall\, k \in \mathcal{K}, \\
& \sum_{k\in\mathcal{K}} \|\boldsymbol{v}_{nk}\|_2^2 \leq P_n^{\max}, \quad \forall\, n \in \mathcal{N}, \\
& |\theta_m| = 1, \quad \forall\, m \in \mathcal{M}.
\end{aligned}
$$

An alternating framework

- **Stage I:** updating beamforming vector $\{\boldsymbol{v}_{nk}\}$ with fixed RIS phase shifts $\boldsymbol{\Theta}$

$$\begin{aligned} \underset{\{\boldsymbol{v}_{nk}\},\boldsymbol{\Theta}}{\text{minimize}} \quad & \sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}}\frac{1}{\eta_n}\|\boldsymbol{v}_{nk}\|_2^2+\sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}}\mathbf{1}_{\{\boldsymbol{v}_{nk}=\mathbf{0}\}}P_{nk}^{\mathrm{c}} \\ \text{subject to} \quad & \mathrm{SINR}_k\geq\gamma_k,\quad\forall\, k, \\ & \sum_{k\in\mathcal{K}}\|\boldsymbol{v}_{nk}\|_2^2\leq P_n^{\max},\quad\forall\, n, \\ & |\theta_m|=1,\quad\forall\, m. \end{aligned}$$

mixed $\ell_{1,2}$-norm for group sparsity inducing

$$\begin{aligned} \underset{\{\boldsymbol{v}_{nk}\}}{\text{minimize}} \quad & \sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}}\frac{1}{\eta_n}\|\boldsymbol{v}_{nk}\|_2^2+\sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}}P_{nk}^{\mathrm{c}}\|\boldsymbol{v}_{nk}\|_2 \\ \text{subject to} \quad & \mathrm{SINR}_k\geq\gamma_k,\quad\forall\, k, \\ & \sum_{k\in\mathcal{K}}\|\boldsymbol{v}_{nk}\|_2^2\leq P_n^{\max},\quad\forall\, n, \\ & |\theta_m|=1,\quad\forall\, m. \end{aligned}$$

An alternating framework

- **Stage II:** updating phase-shift matrix $\boldsymbol{\Theta}$ with fixed beamforming vectors

define $\boldsymbol{a} = [\theta_1, \ldots, \theta_M]^{\mathsf{H}}$, $\boldsymbol{w}_{kl} = \mathrm{diag}(\boldsymbol{h}_{r,k}^{\mathsf{H}})\tilde{\boldsymbol{G}}\boldsymbol{v}_l$, $b_{kl} = \boldsymbol{h}_k^{\mathsf{H}}\boldsymbol{v}_l$, $\boldsymbol{R}_{kl} = \begin{bmatrix} \boldsymbol{w}_{kl}\boldsymbol{w}_{kl}^{\mathsf{H}} & \boldsymbol{w}_{kl}b_{kl}^{\mathsf{H}} \\ \boldsymbol{w}_{kl}^{\mathsf{H}}b_{kl} & 0 \end{bmatrix}$, $\bar{\boldsymbol{a}} = \begin{bmatrix} \boldsymbol{a} \\ t \end{bmatrix}$,

inhomogeneous QCQP

$$\begin{aligned} \text{find} \quad & \boldsymbol{a} \in \mathbb{C}^M \\ \text{subject to} \quad & \frac{|b_{kk} + \boldsymbol{a}^{\mathsf{H}}\boldsymbol{w}_{kk}|^2}{\sum_{l\neq k}|b_{kl} + \boldsymbol{a}^{\mathsf{H}}\boldsymbol{w}_{kl}|^2 + \sigma_k^2} \geq \gamma_k, \forall k, \\ & |a_m|^2 = 1, \forall m. \end{aligned}$$

homogeneous QCQP

$$\begin{aligned} \text{find} \quad & \bar{\boldsymbol{a}} \in \mathbb{C}^{M+1} \\ \text{subject to} \quad & \frac{\bar{\boldsymbol{a}}^{\mathsf{H}}\boldsymbol{R}_{kk}\bar{\boldsymbol{a}} + |b_{kk}|^2}{\sum_{l\neq k}\bar{\boldsymbol{a}}^{\mathsf{H}}\boldsymbol{R}_{kl}\bar{\boldsymbol{a}} + |b_{kl}|^2 + \sigma_k^2} \geq \gamma_k, \forall k, \\ & |\bar{a}_m|^2 = 1, \text{for } m = 1, \ldots, M+1. \end{aligned}$$

matrix lifting $\boldsymbol{A} = \bar{\boldsymbol{a}}\bar{\boldsymbol{a}}^{\mathsf{H}}$

$$\begin{aligned} \text{find} \quad & \boldsymbol{A} \in \mathbb{C}^{(M+1)\times(M+1)} \\ \text{subject to} \quad & \frac{\mathrm{Tr}(\boldsymbol{R}_{kk}\boldsymbol{A}) + |b_{kk}|^2}{\sum_{l\neq k}\mathrm{Tr}(\boldsymbol{R}_{kl}\boldsymbol{A}) + |b_{kl}|^2 + \sigma_k^2} \geq \gamma_k, \forall k, \\ & \boldsymbol{A}_{mm} = 1, \text{for } m = 1, \ldots, M+1, \\ & \boldsymbol{A} \succeq \boldsymbol{0} \ \text{ and } \ \mathrm{rank}(\boldsymbol{A}) = 1. \end{aligned}$$

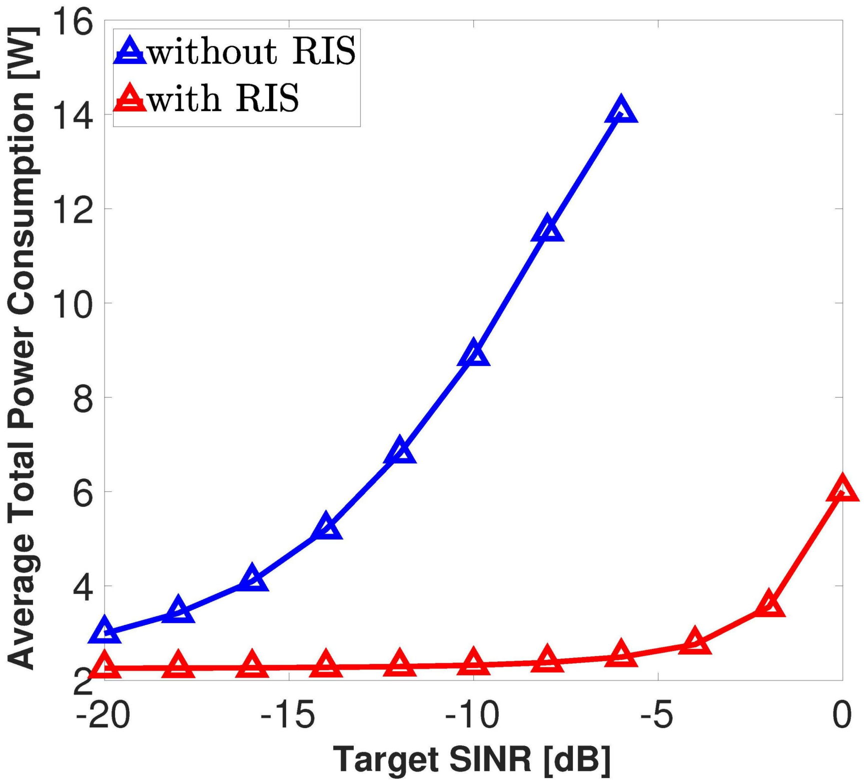
DC representation

$$\mathrm{rank}(\boldsymbol{A}) = 1 \iff \mathrm{Tr}(\boldsymbol{A}) - \|\boldsymbol{A}\|_2 = 0,$$

DC programming

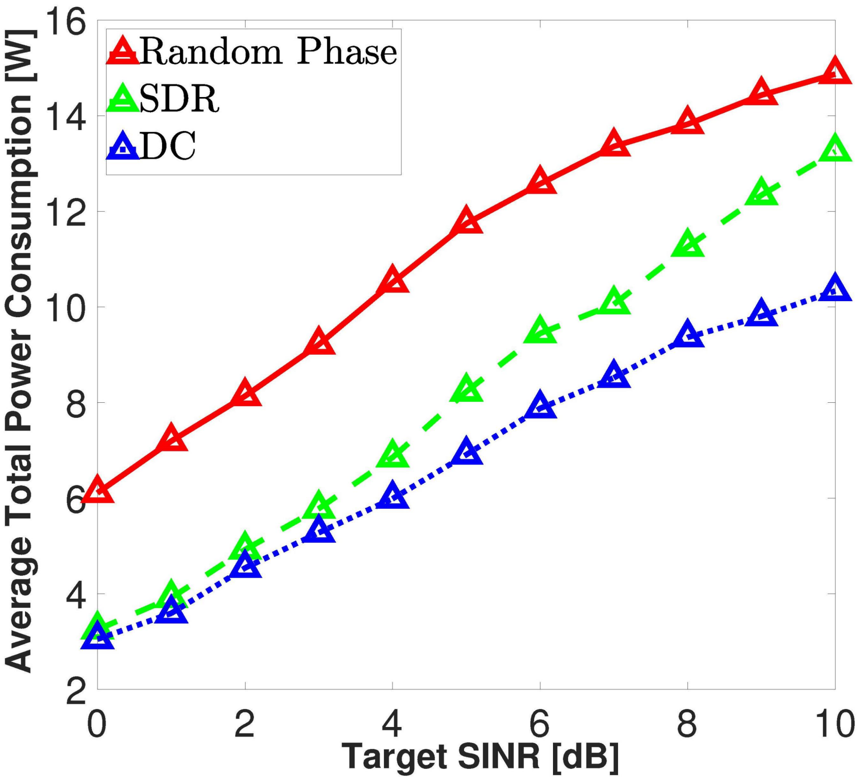
$$\begin{aligned} \underset{\boldsymbol{A}\succeq\boldsymbol{0}}{\text{minimize}} \quad & \mathrm{Tr}(\boldsymbol{A}) - \|\boldsymbol{A}\|_2 \\ \text{subject to} \quad & \frac{\mathrm{Tr}(\boldsymbol{R}_{kk}\boldsymbol{A}) + |b_{kk}|^2}{\sum_{l\neq k}\mathrm{Tr}(\boldsymbol{R}_{kl}\boldsymbol{A}) + |b_{kl}|^2 + \sigma_k^2} \geq \gamma_k, \forall k, \\ & \boldsymbol{A}_{mm} = 1, \text{for } m = 1, \ldots, M+1. \end{aligned}$$

Simulation Results



Insights: deploying an RIS in edge inference system can significantly reduce the total power consumption

Simulation Results



Insights: the proposed DC significantly outperforms two benchmark algorithms in obtaining rank-one solutions

Concluding remarks

- **Edge inference over "intelligent" wireless networks**
  - Edge inference empowered by reconfigurable intelligent surface
- **A mixed $\ell_{1,2}$–norm and DC based alternating framework**
  - Mixed $\ell_{1,2}$-norm for group sparsity inducing
  - DC representation for low-rank functions
  - MM algorithm for DC programming

To learn more…

- **Web:** http://shiyuanming.github.io/publicationstopic.html
- **Papers:**
  - **S. Hua** and Y. Shi, "Reconfigurable intelligent surface for green edge inference in machine learning," in *Proc. IEEE Global Commun. Conf. (Globecom) Workshops*, Waikoloa, Hawaii, USA, Dec. 2019.
  - **S. Hua**, Y. Zhou, K. Yang, and Y. Shi, "Reconfigurable intelligent surface for green edge inference," *submitted to IEEE Trans. Wireless Commun. 2019*, https://arxiv.org/abs/1912.00820.

Thanks

huasheng@shanghaitech.edu.cn