Security Level:

Mathematical and Algorithmic Sciences Lab

Mérouane Debbah
merouane.debbah@huawei.com

www.huawei.com

HUAWEI TECHNOLOGIES CO., LTD.

HUAWEI

Huawei at a Glance

Who is Huawei

- A leading global **ICT** solutions provider
- A **private** company established in 1987
- A **Fortune Global 500** company

Employee
s

- **150,000+** employees worldwide
- **70,000+** engaged in R&D

Market Progress

- **$35.4 B** revenue in 2012(unaudited)
- Serving **45** of the world's top 50 operators
- Serving **1/3** of the world's population

Business Areas

- Carrier network
- Enterprise business
- Consumer devices

Huawei R&D Core Strategy
Customer-centric Innovation

R&D Investment

14% of revenues allocated to R&D

$4.8Bn in R&D investments in 2012

$19Bn R&D investment (2000 – 2012)

70,000+ employees engaged in R&D

Patents

12,453 PCT patent applications

41,948 patent applications in China, and **14,494** patent applications outside of China.

30,240 granted applications, **90%** are invention patents.

Customer-driven R&D system

- IPD process
- Large-scale platform sharing
- CMM5 Quality control systems

16 R&D centers worldwide

28 joint innovation centers with leading operators

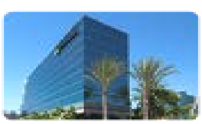
R&D Centers

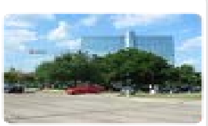
150+ standards organizations

5,000+ standards proposals submitted in 2012

Standards

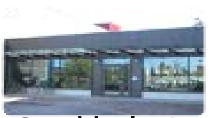
Huawei R&D Centers Worldwide, We Are Open to Whole World

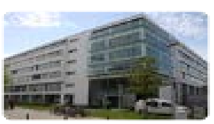

San Diego, USA

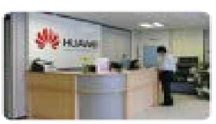

Dallas Texas, USA

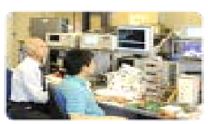

Moscow, Russia


Bangalore, India

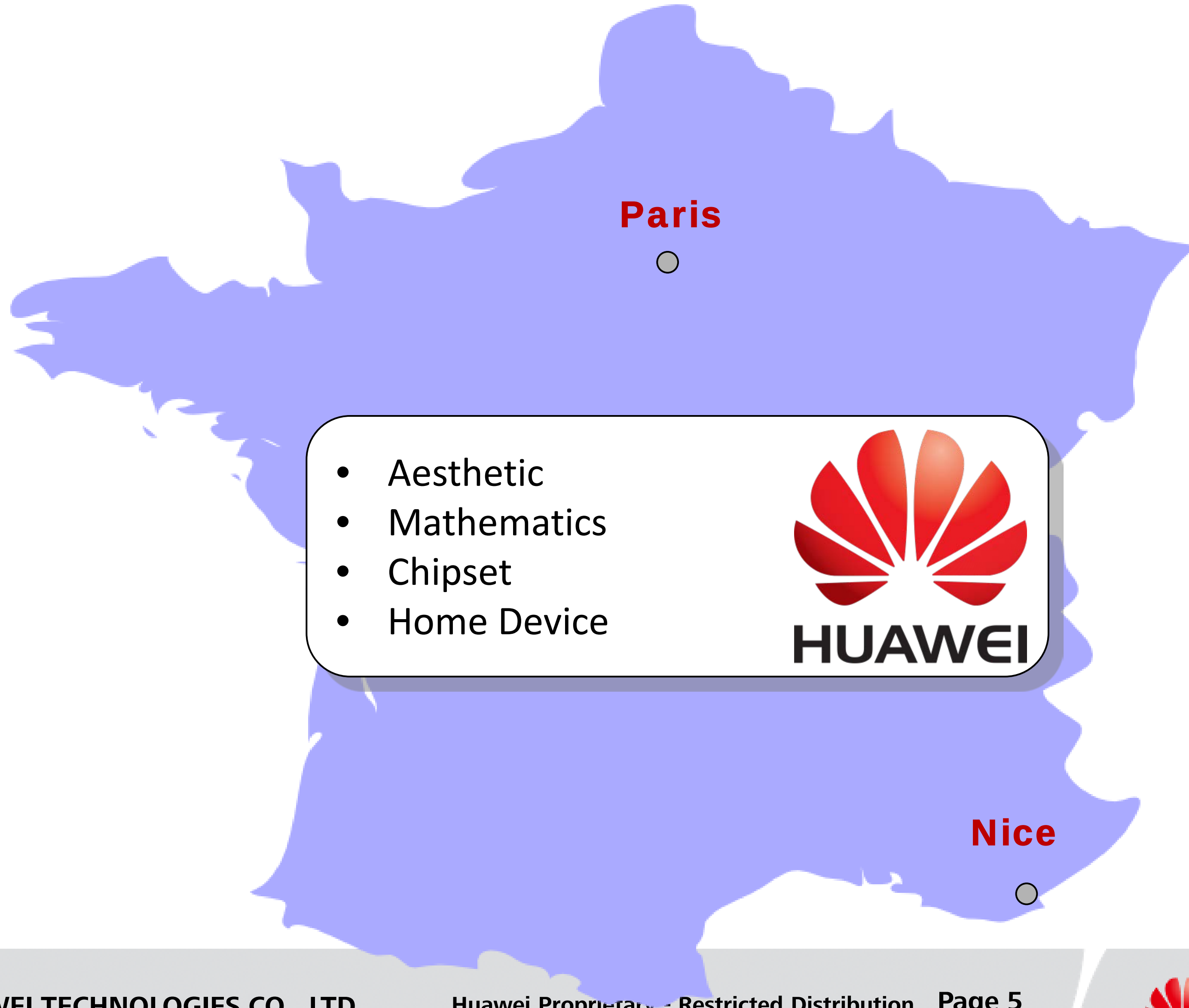

Stockholm/ Gothenburg, Sweden


Munich, Germany

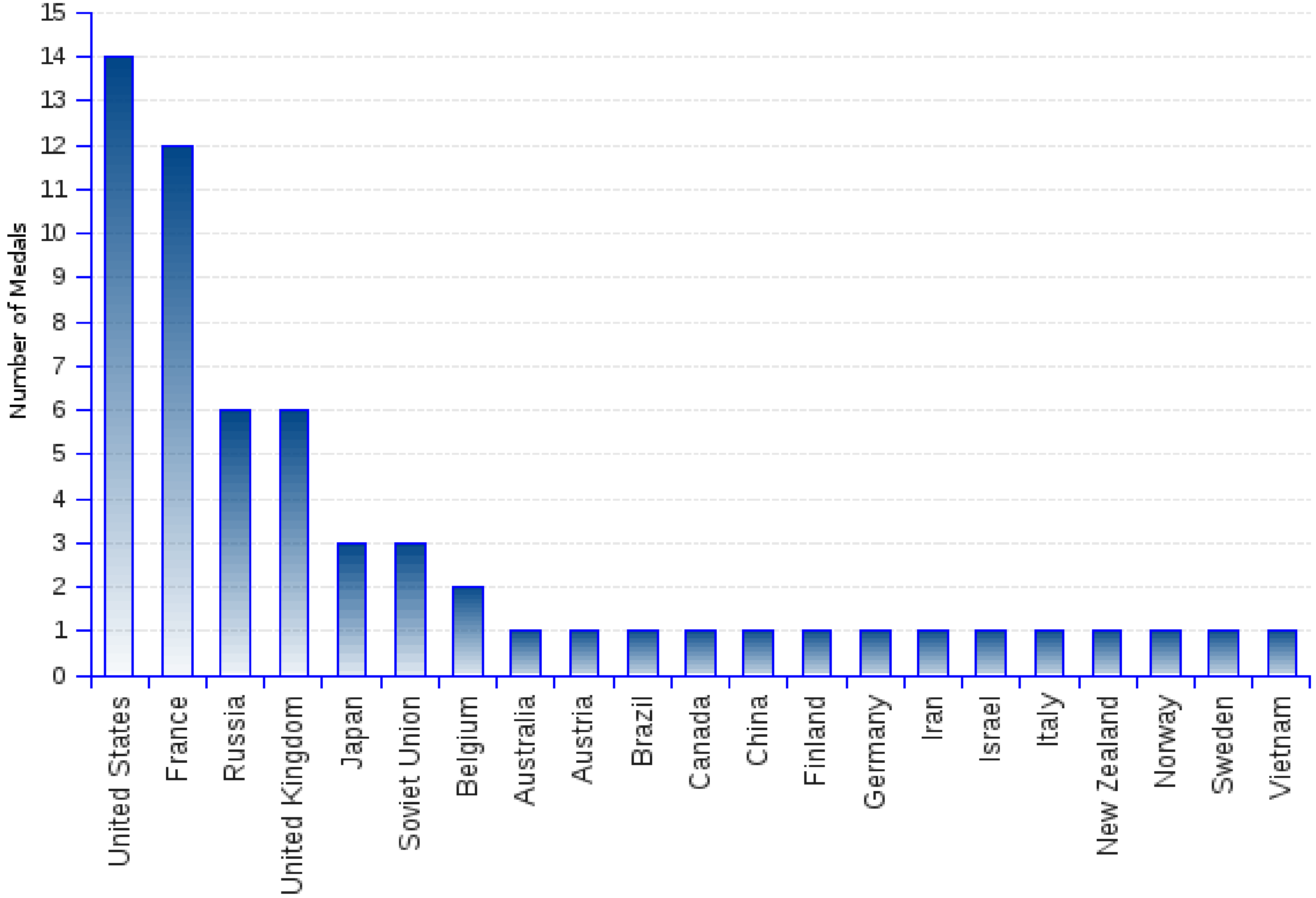

Paris, France


Milan, Italy

Huawei French Research Center



- Aesthetic
- Mathematics
- Chipset
- Home Device


Number of Medals
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
United States
France
Russia
United Kingdom
Japan
Soviet Union
Belgium
Australia
Austria
Brazil
Canada
China
Finland
Germany
Iran
Israel
Italy
New Zealand
Norway
Sweden
Vietnam

Mathematical and Algorithmic Science Lab

- Located in Boulogne Billancourt
- 70 researchers at PhD level
- 3 departments:
  - Content Science,
  - Communication Science
  - Networking Science
- Focus: 4.5G, 5G, beyond 5G, compression, and SDN

Content Science Department

- **Cryptography and Data Security** : cryptography, Information Theory, security protocols
- **Compression and Source Coding :** Feedback issues, broadcast channel, cooperation, network coding, compression in network, distributed source coding.
- **Coding for Data Networks:** distributed storage, regenerating codes, network coding, network information theory

Communication Science Department

- **Architecture Modeling and Performance Analysis :** stochastic geometry, distributed optimization, continuous optimization, continuous game theory, extreme value theory, machine learning, multi-armed bandits
- **Signal and Information Processing :** Channel estimation and modeling, multi-user communication theory, low-complexity physical layer techniques, optimization, distributed methods, random matrices, Bayesian inference
- **Communication Algorithms Design :** transceive algorithmic design, modulation theory, multiple access techniques, resource allocation algorithms, multiantenna systems design, cognitive and cooperative communications, compressed sensing
- **Optical and Physics :** Optical Communications, Quantum communications, Replica Analysis, Free Probability Theory, non-linear medium, Mean-field theory, cavity methods

Networking Science Department

- **Computational Learning, Algorithms and Networks :** Machine learning, network science, network economics, game theory, community detection/clustering, dynamics and distributed algorithms on networks
- **Network Control and Resource Allocation :** Resource Allocation, Dynamic Control, Stability, Network Algorithms, Performance analysis, Caching , congestion control, networking big data, multi-hop networks, game theory
- **Network and Traffic Optimization :** Combinatorial algorithms, path computation, flow allocation, linear and non-linear programming, constraint programming, parallel algorithms, network virtualization, function placement.
- **Context Aware Optimization :** Anticipatory adaptation, user-centric resource allocation, proactive buffering, cross-layer design, channel and traffic models, learning and prediction

5G to Enable a Connected World

HUAWEI

5G From Mobile Internet to Connecte

Voice

Smartphone

3G

4G

Mobile Internet

(4 Billions@2020)

Mobile Internet replaced PC Internet

HD Video Surveillance

Augmented Reality

Wireless IP TV

High Speed Train

Meter & Sensor

Stadium

4K 3D HD TV

Smartphone

Voice

Shipping Logistic

Multi-User UHD Telepresence

Gaming

Wireless Cloud Office

Automatic Driving

Monitoring

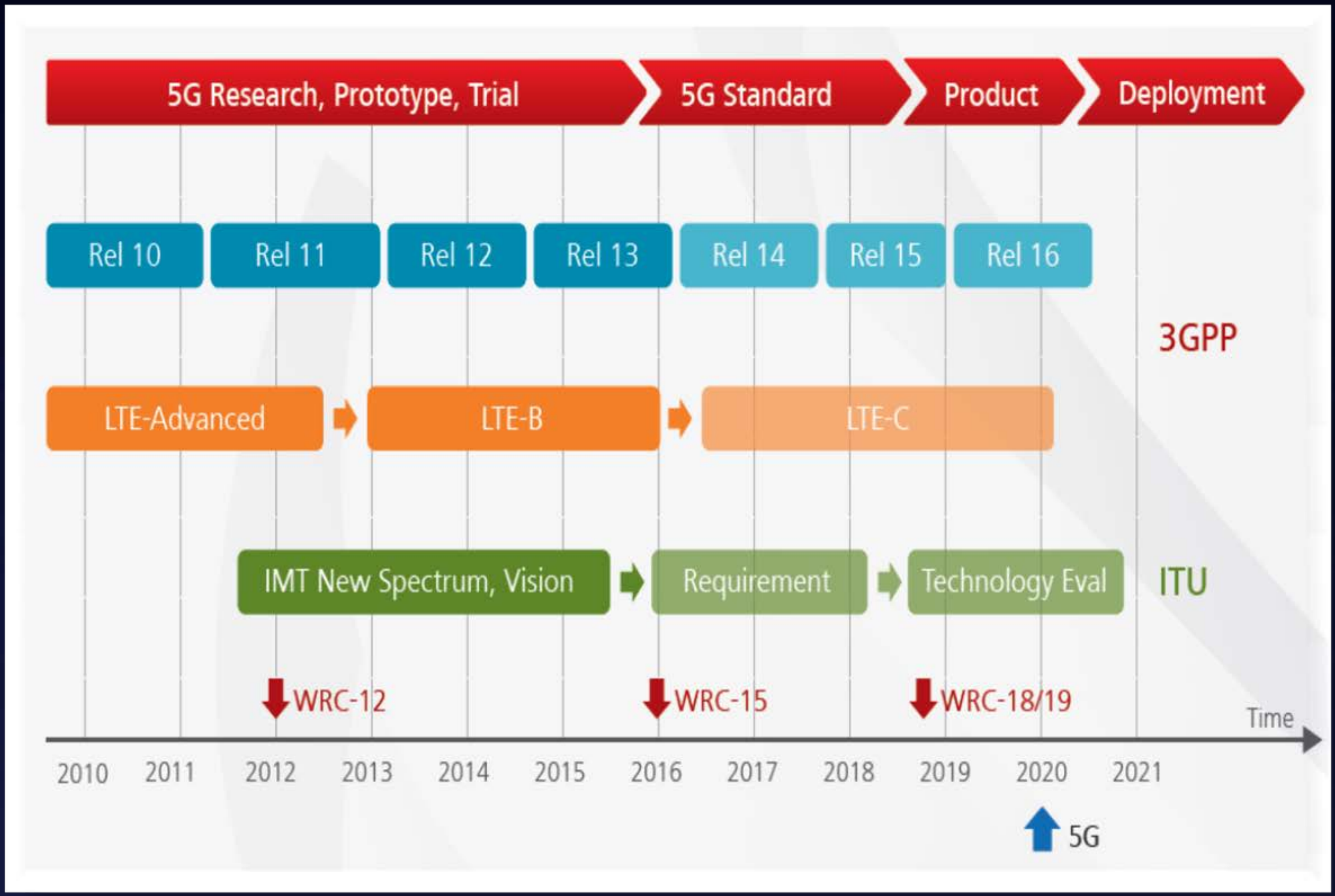
5G

Connected World

(50-100Billions@2020)

90% objects are not connected

5G Timeline



5G Research, Prototype, Trial
5G Standard
Product
Deployment
Rel 10
Rel 11
Rel 12
Rel 13
Rel 14
Rel 15
Rel 16
3GPP
LTE-Advanced
LTE-B
LTE-C
IMT New Spectrum, Vision
Requirement
Technology Eval
ITU
WRC-12
WRC-15
WRC-18/19
Time
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
5G

Example: movie projectors tomorrow (lasers)

➔ ***30-50 Mb/s* for a single view transmission and *Zero-Latency* (adaptive) interaction client-server ***

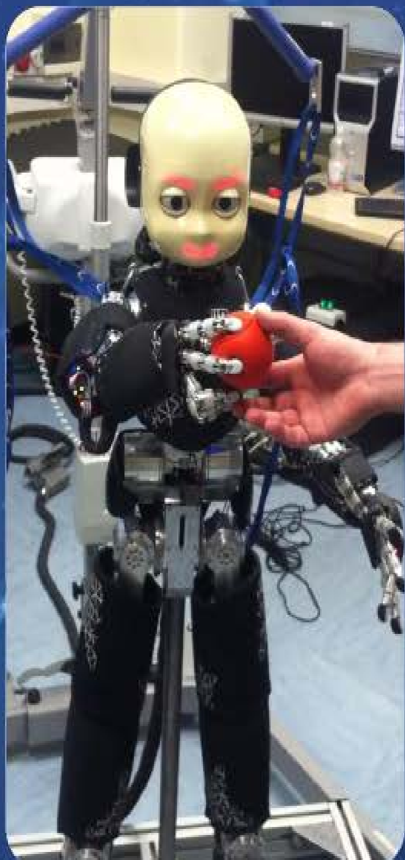
**) For luminance (brightness), chrominance (color), resolution, view point, etc. adaptation*

2-8K ➔ 30-50 Mb/s/view

http://spectrum.ieee.org/consumer-electronics/audiovideo/lasers-coming-to-a-theater-near-you

Example: The iCub robot platform (www.iit.it)

→ ***5.000 sensors!***



iit, Genova, Nov 2014

Computer vision

Force control

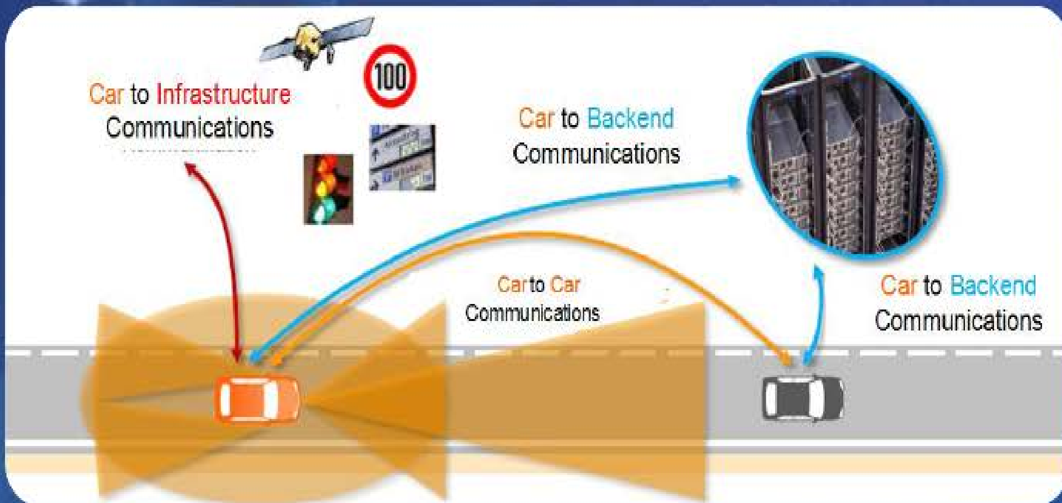
Sensor	Specs	Bandwidth
Cameras	2x, 640x480, 30fps, 8/24bit	147Mbit/s uncompressed
Microphones	2x, 44kHz, 16bit	1.4Mbit/s
F/T sensors	6x, 1kHz, 8bit	48kbit/s
Gyroscopes	12x, 100Hz, 16bit	19.2kbit/s
Tactile sensors	4000x, 50Hz, 8bit	1.6Mbit/s
Control commands	53DoF x 2-4 commands, 100Hz/1kHz, 16bit	3.3Mbit/s (worst case), 170kbit/s (typical)

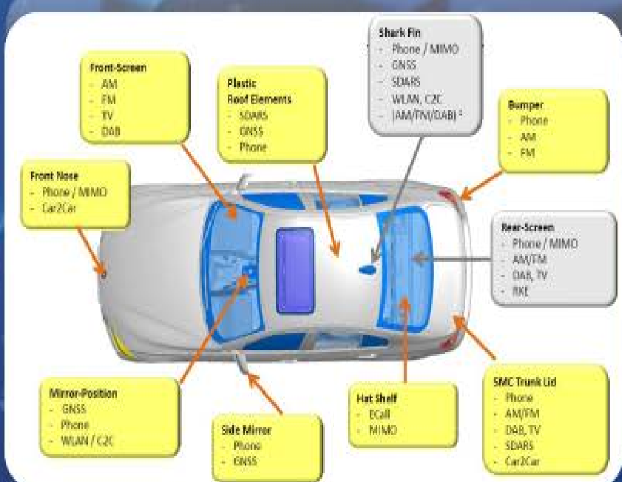
→ *Force control latency requirement = 1-5 ms*

[G. Metta "Robotics-Derived Requirements for the Internet of Things in the 5G Context," IEEE MMTC E-Letter, Sept 2014]

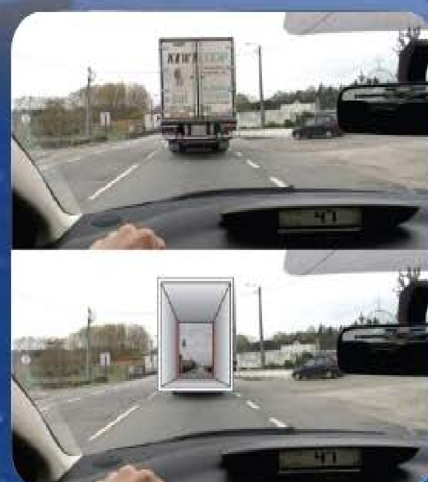
Example: Future Car Communications

➔ ***New Antenna Concepts for MIMO, Integration of 11p and LTE/5G, Mobile Edge Computing***





[Kathrein Automotive]



[Markus Dillinger, Huawei]

Communication requirements

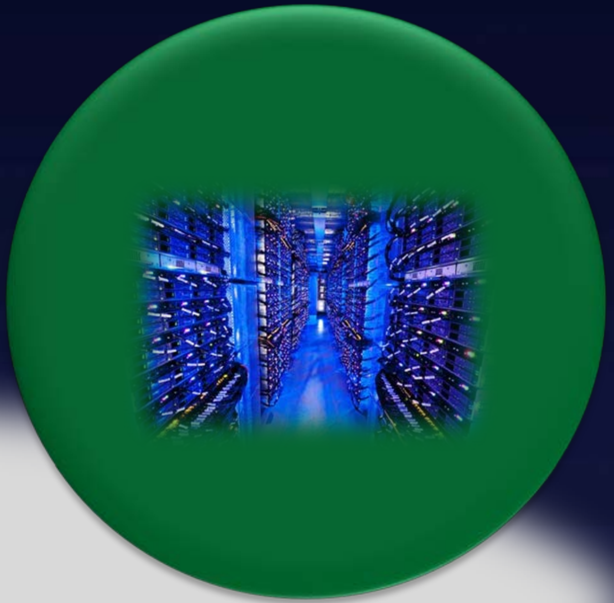
- Better connection than smart phone
- Reliable for future advanced driver assistant systems (ADAS)
- High data volumes (>200MB/s) at low latencies for future cooperative automatic driving functions (V2V)
- Support performance up to maximum speed (500km/h relative)
- Any network operator, regardless vehicle occupants' contract (safety information)

5G World

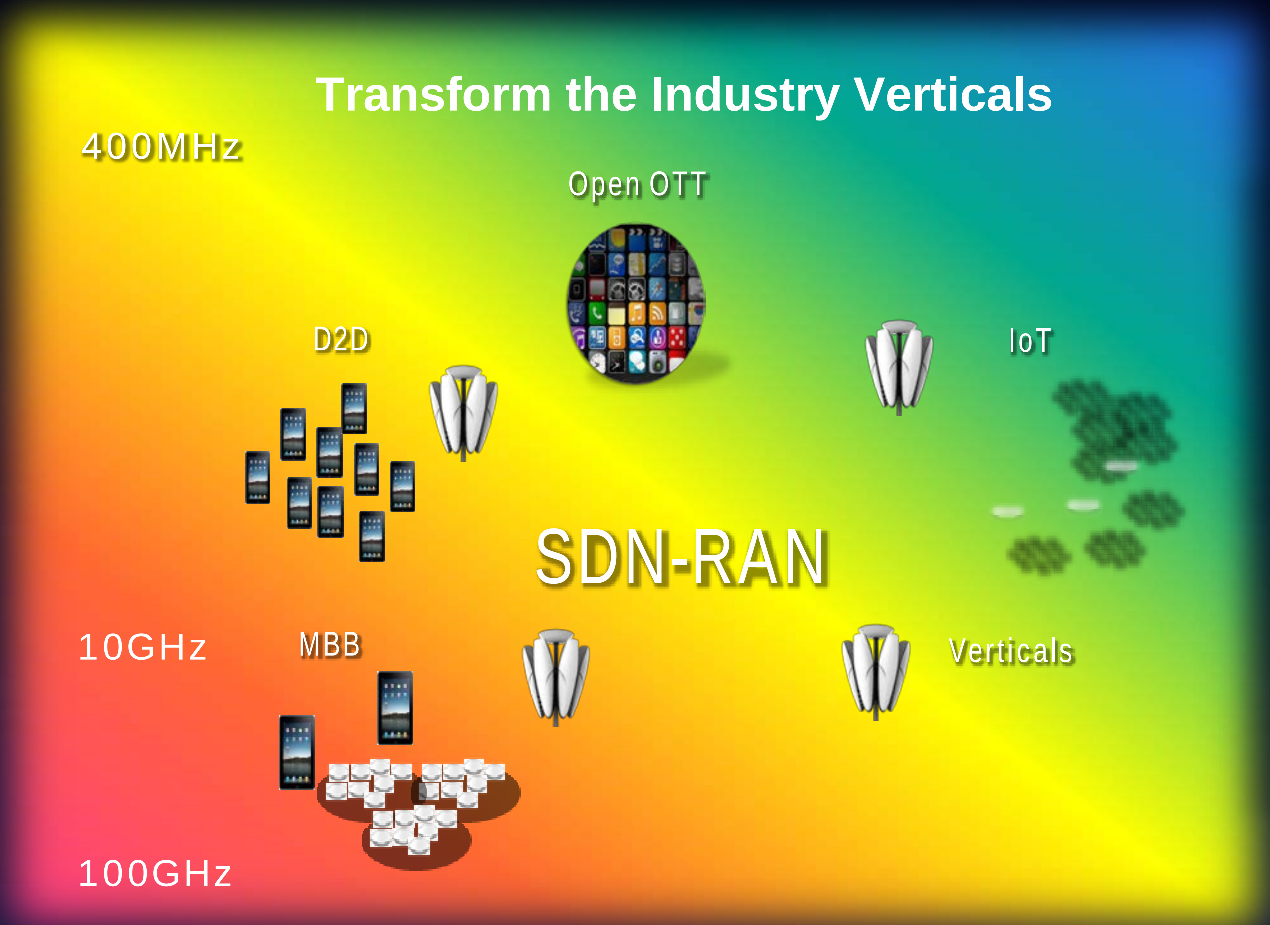
Everything on Mobile

Everything Connected

Everything Virtualized



5G (Beyond Smartphone)



Auto-drive

Medicare

Robots

Meters Sensors

Capacity
1000X
（Capacity/k

Speed
100X
（10Gbps)

Latency
Less than 1ms

Links
100x

Energy
1000X
Reduce

5G Expands into Diverse Domains

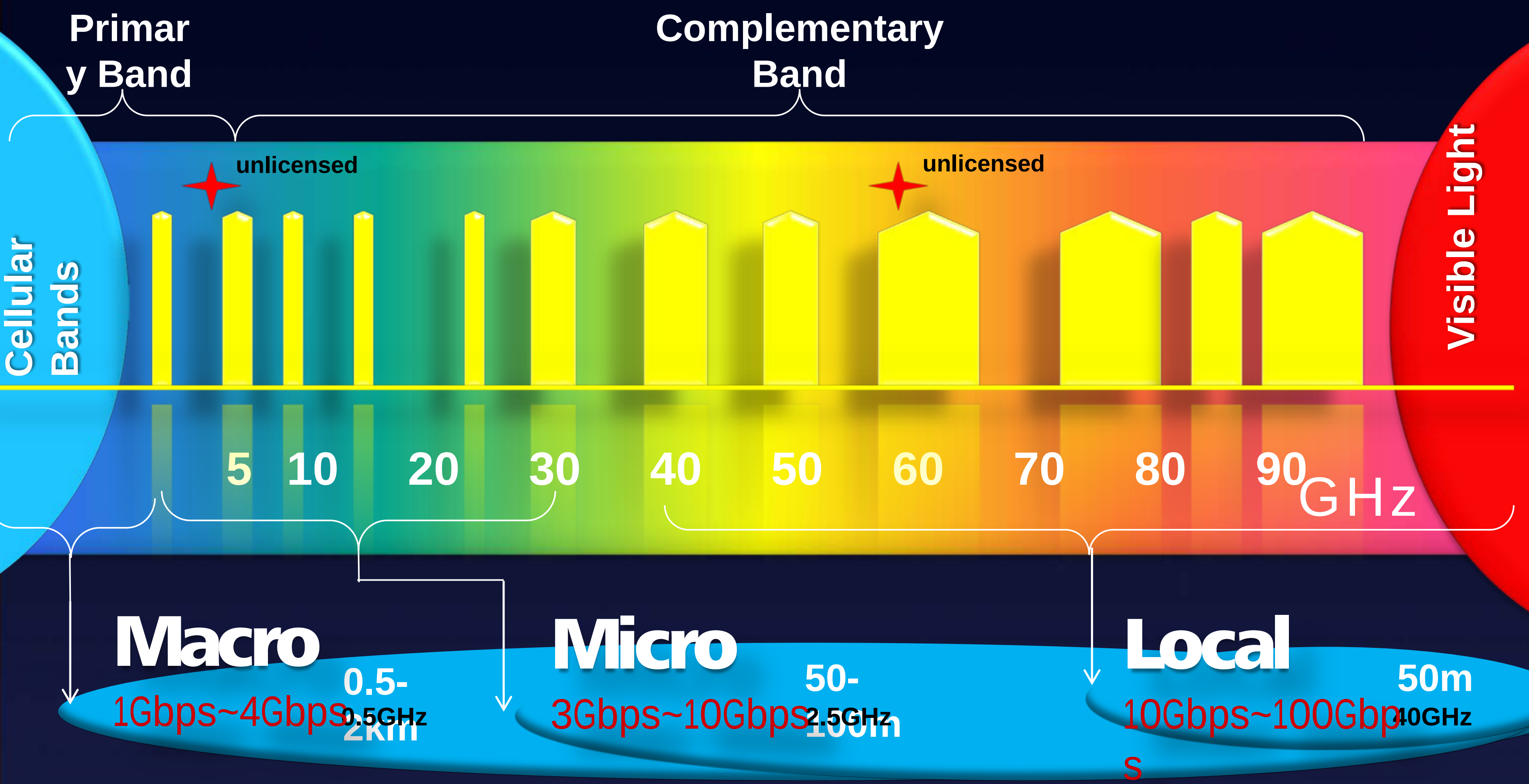
5G is Cellular MBB

5G is Wireless IoT

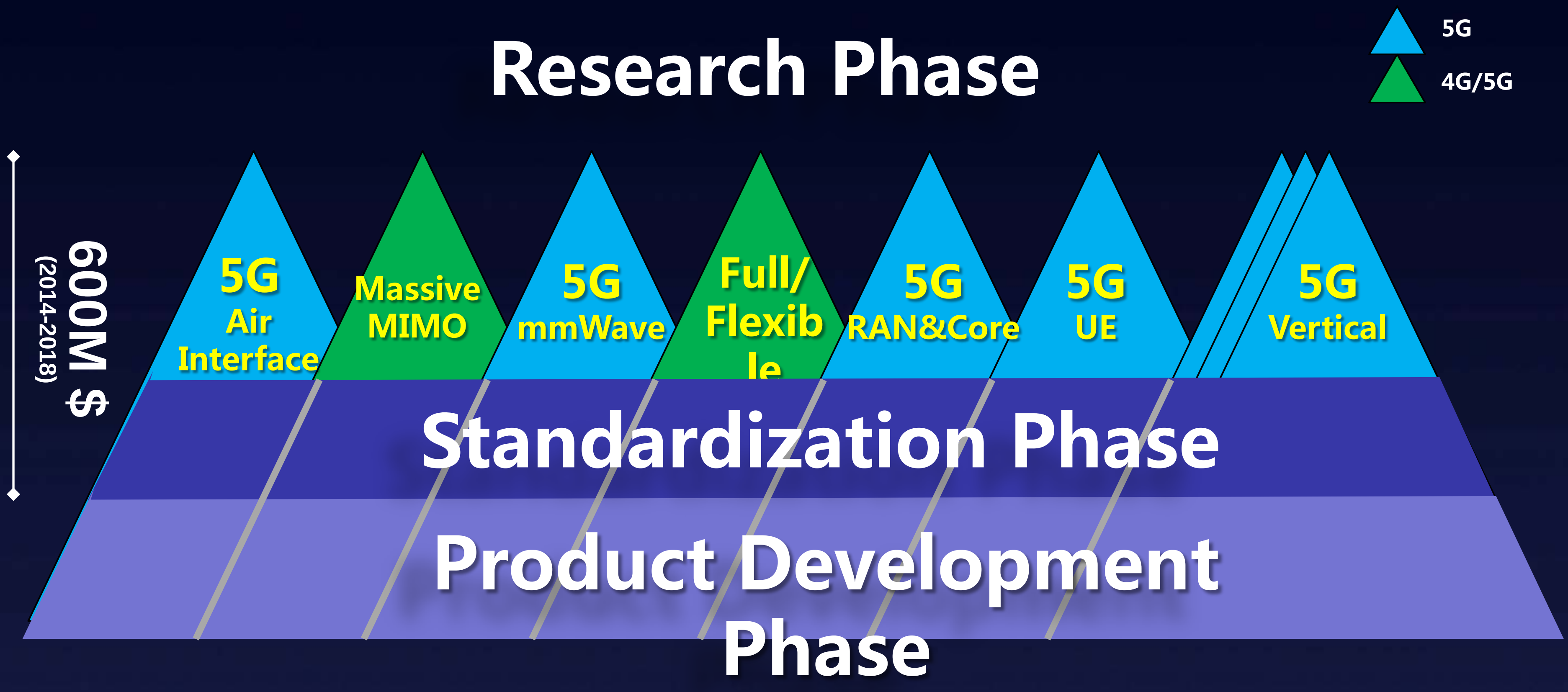
5G is mmWave UDN

5G is NfV and SDN

Single & Unified Air-Interface for All Spectrum Access



Foundational Technologies



The New Landscape of RAN in 5G

NfV/SDN is the Birth-bed for 5G

User centric mobility instead of network centric

No-Cell RAN and De-Cell-lization

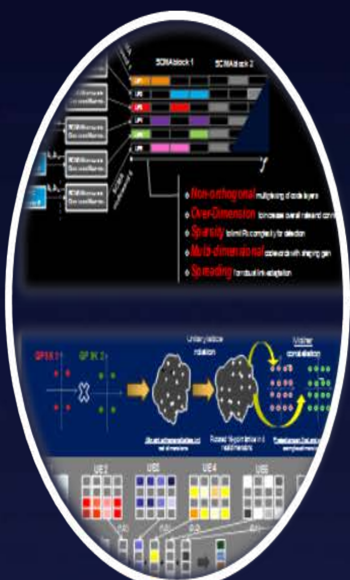
Soft Defined Protocol

Soft Defined RAN

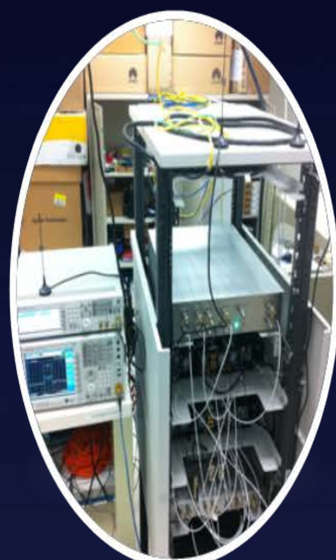
Soft Defined Network Topology

Huawei 5G Technology Breakthroughs

SCMA Air-Interface

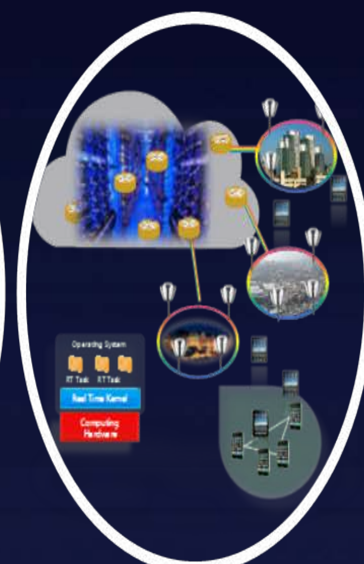


Full-Duplex



50Gbps Basest ion

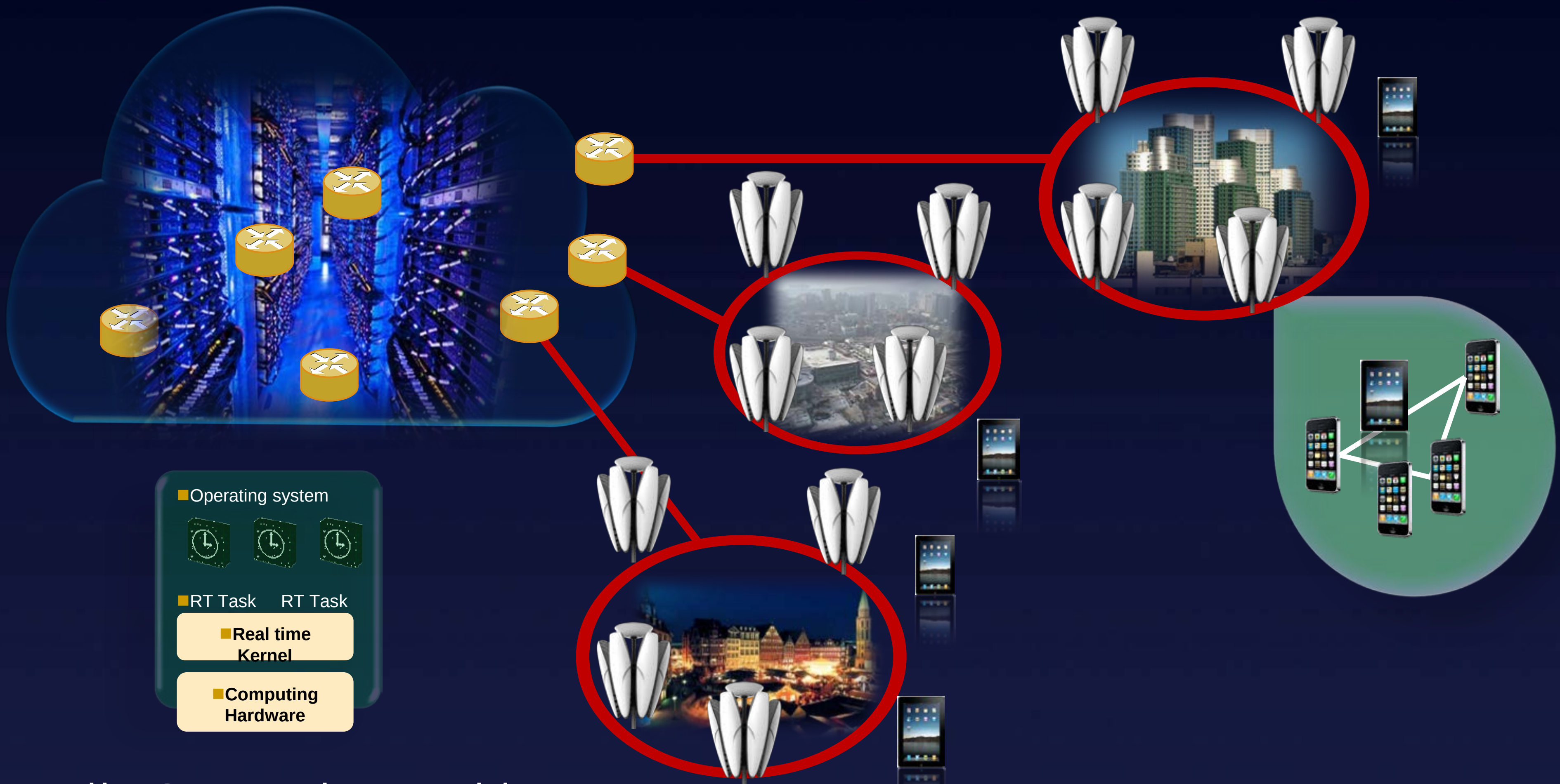
Virtualized Radio



100Gbps mmWave

C-RAN Cloud Radio Access Networks

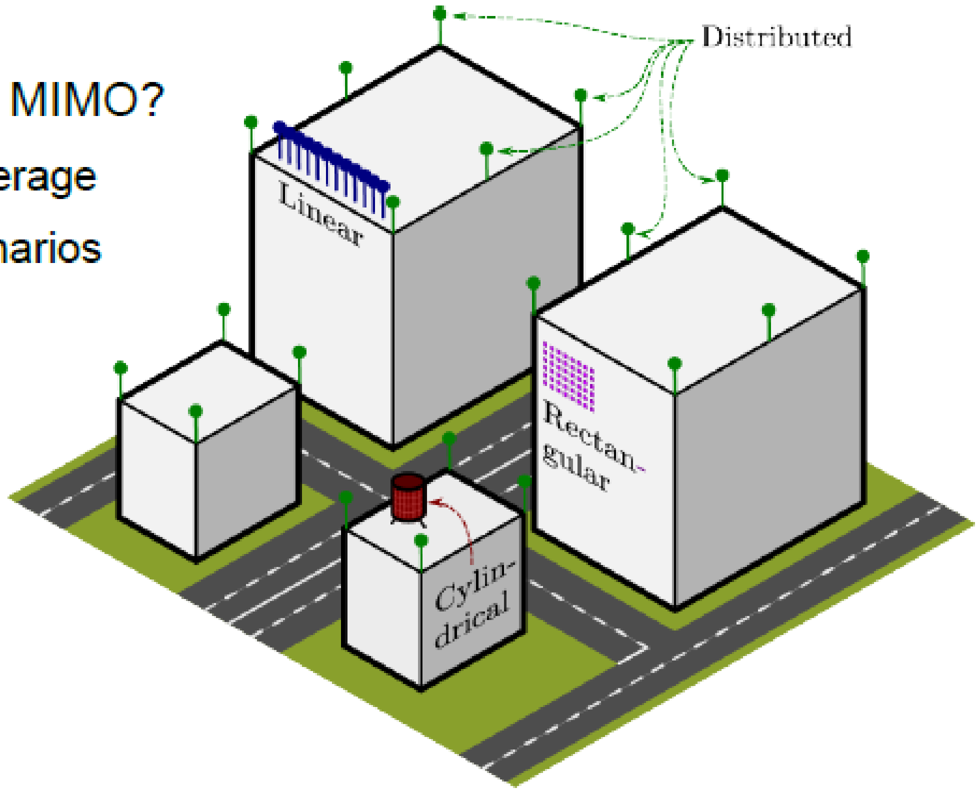
Huawei Innovation:
Hyper-Transceiver Architecture



Radio Computing Architecture

Massive MIMO Deployment

- When to Deploy Massive MIMO?
  - Improved wide-area coverage
  - Special superdense scenarios
- Co-located Deployment
  - 1D, 2D, or 3D arrays
  - One or multiple sectors
- Distributed Deployment
  - Remote radio heads
  - Cloud RAN



Topics of Interest

Clean Slate Design of 5G

Where to start from?

- **Tons of Plenary Talks and Overview Articles**
  - Fulfilling dream of ubiquitous wireless connectivity

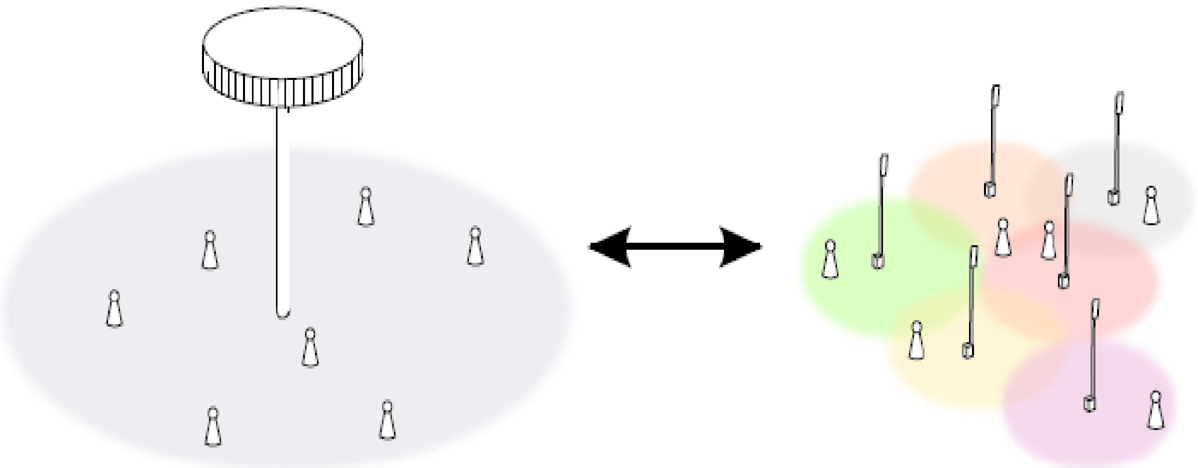
- **Expectation: Many Metrics Should Be Improved in 5G**
  - Higher user data rates
  - Higher area throughput
  - Great scalability in number of connected devices
  - Higher reliability and lower latency
  - Better coverage with more uniform user rates
  - Improved energy efficiency

- **These are Conflicting Metrics!**

- Higher user data rate

The clean slate approach

"David vs Goliath" or "Small Cells vs Massive MIMO"



How to densify: *"More antennas or more BSs?"*

What if we are only interested in the average throughput per UT?

A thought experiment

Consider an infinite large network of randomly uniformly distributed base stations and user terminals.

What would be better?

A $2\times$ more base stations

B $2\times$ more antennas per base station

How to optimally deploy your antennas?

A thought experiment

Consider an infinite large network of randomly uniformly distributed base stations and user terminals.

What would be better?

A $2 \times$ more base stations

B $2 \times$ more antennas per base station

Stochastic geometry can provide an answer.

What if we are only interested in the average throughput per UT?

System model: Downlink

Received signal at a tagged UT at the origin:

$$\boxed{y = \underbrace{\frac{1}{r_0^{\alpha/2}} \mathbf{h}_0^{\mathsf{H}} \mathbf{x}_0}_{\text{desired signal}} + \underbrace{\sum_{i=1}^{\infty} \frac{1}{r_i^{\alpha/2}} \mathbf{h}_i^{\mathsf{H}} \mathbf{x}_i}_{\text{interference}} + n}$$

- $\mathbf{h}_i \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_N)$: fast fading channel vectors
- r_i: distance to ith closest BS
- $P = \mathbb{E}\left[\mathbf{x}_i^{\mathsf{H}} \mathbf{x}_i\right]$: average transmit power constraint per BS

Assumptions:

- infinitely large network of uniformly randomly distributed BSs and UTs with densities λ_{BS} and λ_{UT}, respectively
- single-antenna UTs, N antennas per BS
- each UT is served by its *closest* BS
- distance-based path loss model with path loss exponent $\alpha > 2$
- total bandwidth W, re-used in each cell

What if we are only interested in the average throughput per UT?

Transmission strategy: Zero-forcing

Assumptions:

- $\mathcal{K} = \frac{\lambda_{\mathrm{UT}}}{\lambda_{\mathrm{BS}}}$ UTs need to be served by each BS on average
- total bandwidth W divided into $L \geq 1$ sub-bands
- $K = \mathcal{K}/L \leq N$ UTs are simultaneously served on each sub-band

Transmit vector of BS i:

$$\mathbf{x}_i = \sqrt{\frac{P}{K}} \sum_{k=1}^{K} \mathbf{w}_{i,k} s_{i,k}$$

- $s_{i,k} \sim \mathcal{CN}(0,1)$: message determined for UT k from BS i
- $\mathbf{w}_{i,k} \in \mathbb{C}^{N \times 1}$: ZF-beamforming vectors

What if we are only interested in the average throughput per UT?

Performance metric: Average throughput

Received SINR at tagged UT:

$$\gamma = \frac{r_0^{-\alpha} \left|\mathbf{h}_0^{\mathrm{H}} \mathbf{w}_{0,1}\right|^2}{\sum_{i=1}^{\infty} r_i^{-\alpha} \sum_{k=1}^{K} \left|\mathbf{h}_i^{\mathrm{H}} \mathbf{w}_{i,k}\right|^2 + \frac{K}{P}} = \frac{r_0^{-\alpha} S}{\sum_{i=1}^{\infty} r_i^{-\alpha} g_i + \frac{K}{P}}$$

Coverage probability:

$$P_{\mathrm{cov}}(T) = \mathbb{P}(\gamma \geq T)$$

Average throughput per UT:

$$\boxed{C = \frac{W}{L} \times \mathbb{E}\left[\log(1+\gamma)\right] = \frac{W}{L} \times \int_0^{\infty} P_{\mathrm{cov}}\left(e^z - 1\right) dz}$$

Remarks:

- expectation with respect to fading *and* BSs locations
- $S = \left|\mathbf{h}_0^{\mathrm{H}} \mathbf{w}_{0,1}\right|^2 \sim \Gamma(N-K+1,1)$, $\quad g_i = \sum_{k=1}^{K} \left|\mathbf{h}_i^{\mathrm{H}} \mathbf{w}_{i,k}\right|^2 \sim \Gamma(K,1)$
- K impacts the interference distribution, N impacts the desired signal
- for $P \to \infty$, the SINR becomes independent of λ_{BS}

What if we are only interested in the average throughput per UT?

A closed-form result

Theorem (Combination of Baccelli'09, Andrews'10)

$$P_{cov}(T) = \int_{r_0>0}\int_{-\infty}^{\infty} \mathcal{L}_{I_{r_0}}\left(i2\pi r_0^{\alpha} Ts\right)\exp\left(-\frac{i2\pi r_0^{\alpha} TK}{P}s\right)\frac{\mathcal{L}_S\left(-i2\pi s\right)-1}{i2\pi s} f_{r_0}(r_0)\,ds\,dr_0$$

where

$$\mathcal{L}_{I_{r_0}}(s) = \exp\left(-2\pi\lambda_{BS}\int_{r_0}^{\infty}\left(1-\frac{1}{(1+sv^{-\alpha})^K}\right)v\,dv\right)$$

$$\mathcal{L}_S(s) = \left(\frac{1}{1+s}\right)^{N-K+1}$$

$$f_{r_0}(r_0) = 2\pi\lambda_{BS} r_0 e^{-\lambda_{BS}\pi r_0^2}$$

The computation of $P_{cov}(T)$ requires in general three numerical integrals.

J. G. Andrews, F. Baccelli, R. K. Ganti, "A Tractable Approach to Coverage and Rate in Cellular Networks" IEEE Trans. Wireless Commun., submitted 2010.

F. Baccelli, B. Błaszczyszyn, P. Mühlethaler, "Stochastic Analysis of Spatial and Opportunistic Aloha" Journal on Selected Areas in Communications, 2009

What if we are only interested in the average throughput per UT?

Example

- Density of UTs: $\lambda_{\mathrm{UT}} = 16$
- Constant transmit power density: $P \times \lambda_{\mathrm{BS}} = 10$
- Number of BS-antennas: $N = \lambda_{\mathrm{UT}} / \lambda_{\mathrm{BS}}$
- Path loss exponent: $\alpha = 4$
- UT simultaneously served on each band: $K = \lambda_{UT} / (\lambda_{\mathrm{BS}} \times L)$

$\Rightarrow$ Only two parameters: λ_{BS} and L

Table: Average spectral efficiency C/W in (bits/s/Hz)

sub-bands L	$\lambda_{\mathrm{BS}} = 1$	$\lambda_{\mathrm{BS}} = 2$	$\lambda_{\mathrm{BS}} = 4$	$\lambda_{\mathrm{BS}} = 8$	$\lambda_{\mathrm{BS}} = 16$
1	0.6209	0.8188	1.1964	1.5215	2.1456
2	1.1723	1.2414	1.3404	1.5068	x
4	0.8882	0.8973	1.1964	x	x
8	0.5689	0.5952		x	x
16	0.3532	x	x	x	x

Fully distributing the antennas gives highest throughput gains!

Let us know focus on two metrics…

- **Expectation: Many Metrics Should Be Improved in 5G**
  - Higher user data rates
  - Higher area throughput
  - Great scalability in number of connected devices
  - Higher reliability and lower latency
  - **Better coverage with more uniform user rates**
  - **Improved energy efficiency**

- **These are Conflicting Metrics!**
  - Difficult to maximize theoretically all metrics simultaneously
  - **Our goal: High energy efficiency (EE) with uniform user rates**

How to Measure Energy-Efficiency?

- Energy-Efficiency (EE) in bit/Joule

$$EE = \frac{\text{Average Sum Rate [bit/s/cell]}}{\text{Power Consumption [Joule/s/cell]}}$$

- Conventional Academic Approaches:
  - Maximize rates with fixed power
  - Minimize transmit power for fixed rates

New Problem: Balance rates and power consumption

Important to account for overhead signaling and circuit power!

Optimal Multi-Cell System Design: ZF Beamforming

Optimum

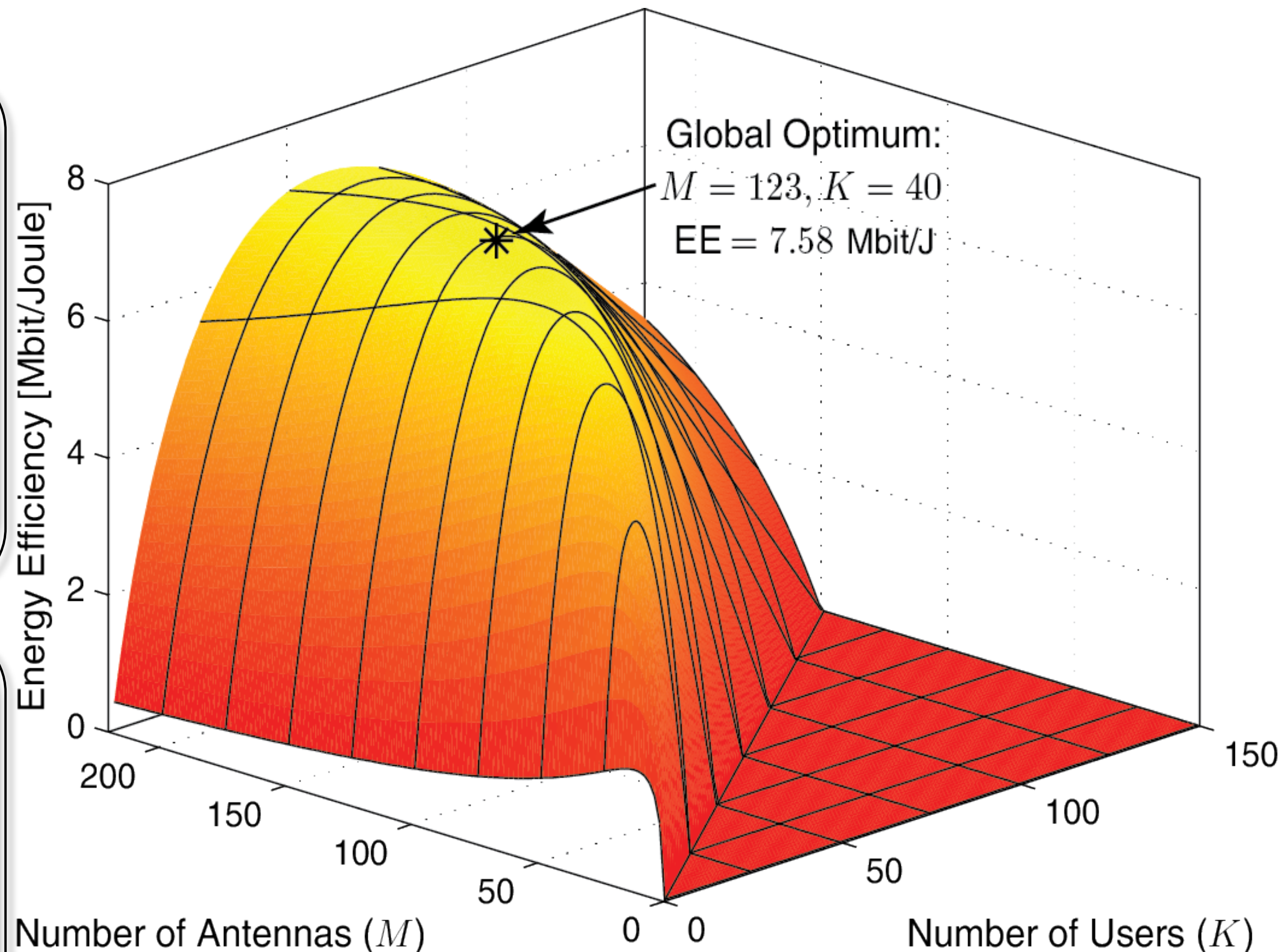
$M = 123$
$K = 40$
$\alpha = 0.28$
$\tau^{(\mathrm{ul})} = 4$

User rates:
≈4-QAM

Massive MIMO!

Many BS antennas

Note that
$M/K \approx 3$



Optimize more than Energy-Efficiency

- **Recall: Many Metrics in 5G Discussions**
  - Average rate (Mbit/s/active user)
  - Average area rate (Mbit/s/km^2)
  - Energy-efficiency (Mbit/Joule)
  - Active devices (per km^2)
  - Delay constraints (ms)

Communications

- **So Far: Only cared about EE**
  - Ignored all other metrics

Optimize Multiple Metrics

We want efficient operation w.r.t. all objectives

Is this possible?
For all at the same time?

Multi-Objective Network Optimization

Example: Results

3 Objectives

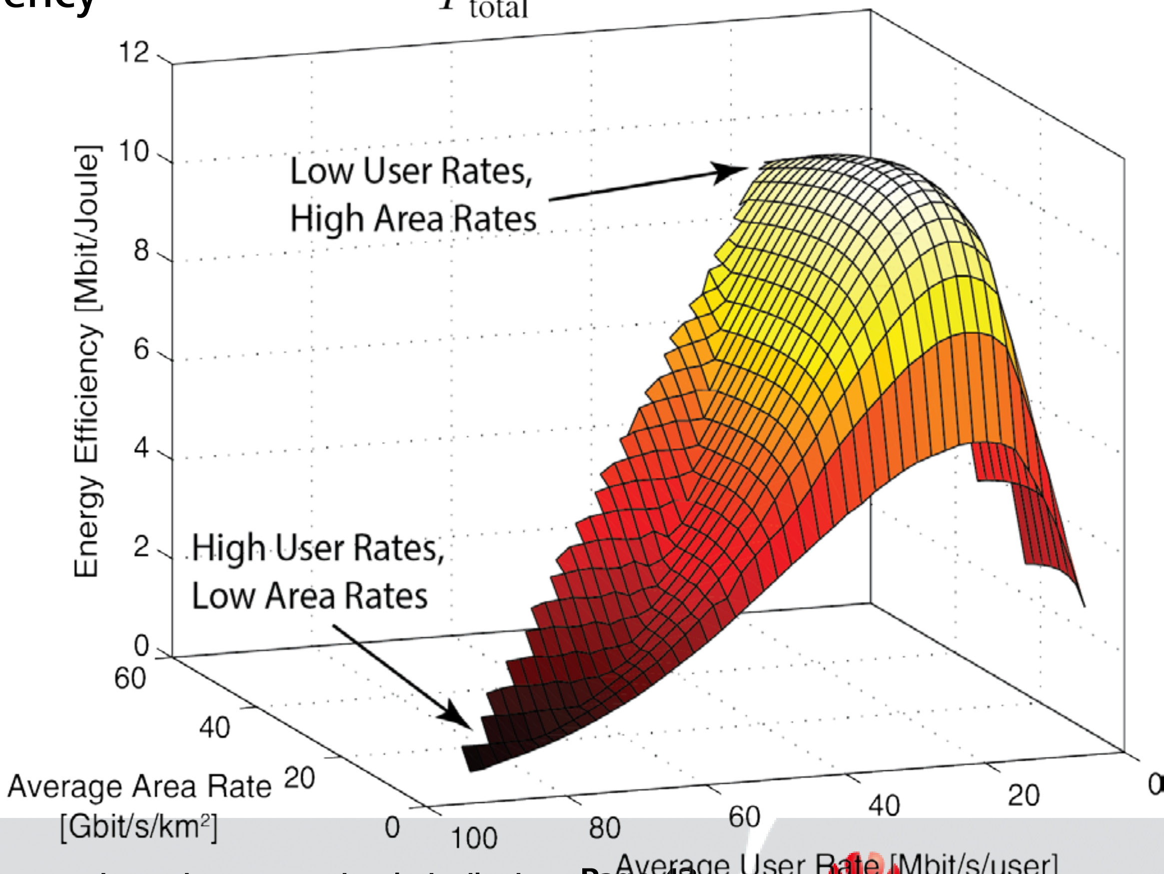
1. Average user rate $g_1(\mathbf{x}) = R_{\text{average}}$ [bit/s/user]
2. Total area rate $g_2(\mathbf{x}) = \frac{K}{A} R_{\text{average}}$ $[\text{bit/s/km}^2]$
3. Energy-efficiency $g_3(\mathbf{x}) = \frac{K R_{\text{average}}}{P_{\text{total}}}$ [bit/J]

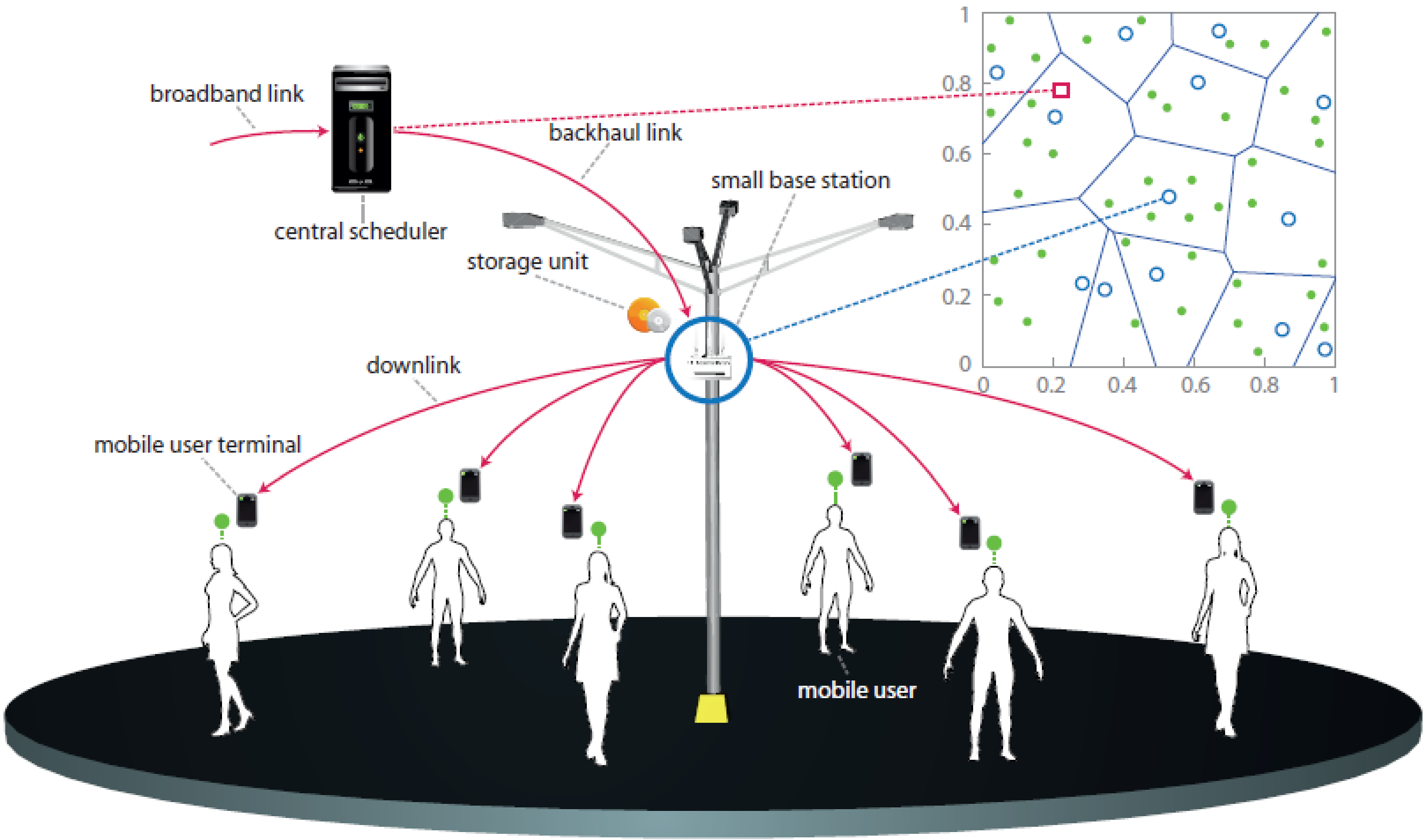
Observations

Area and user rates are conflicting objectives

Only energy efficient at high area rates

Different number of users

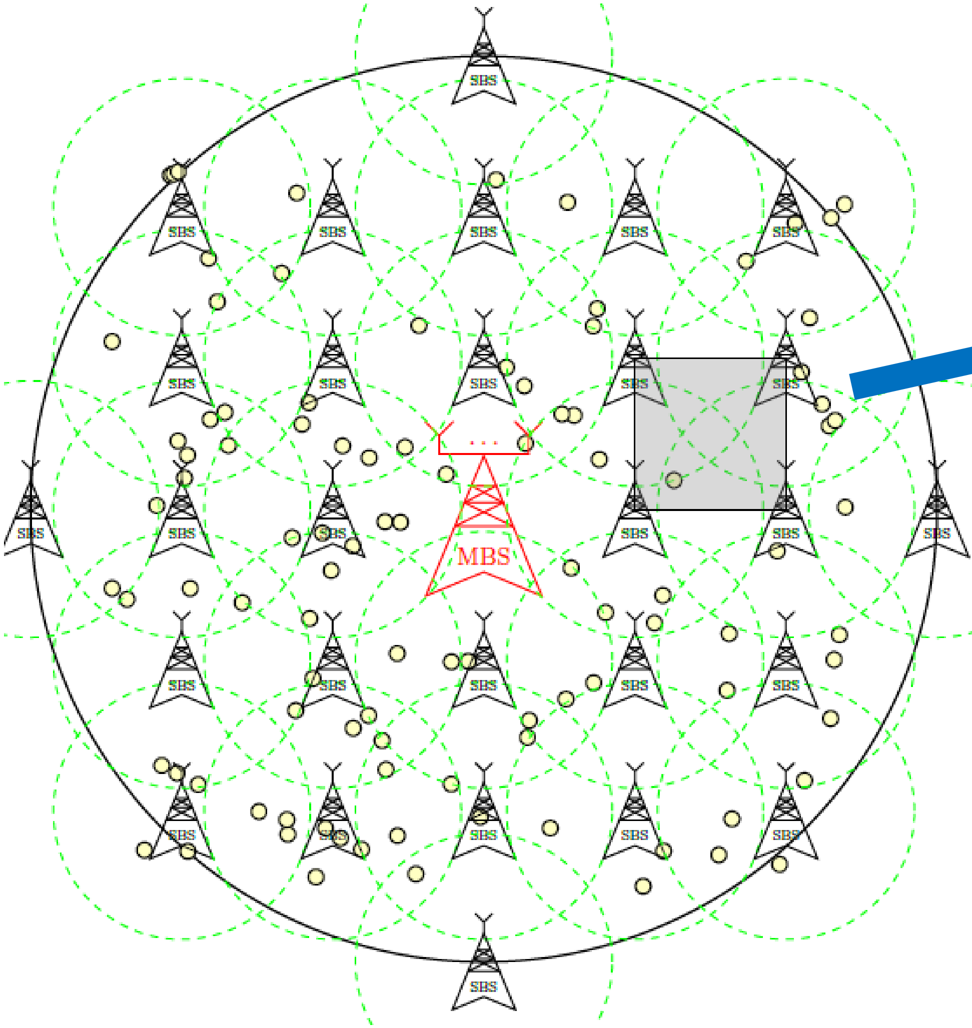


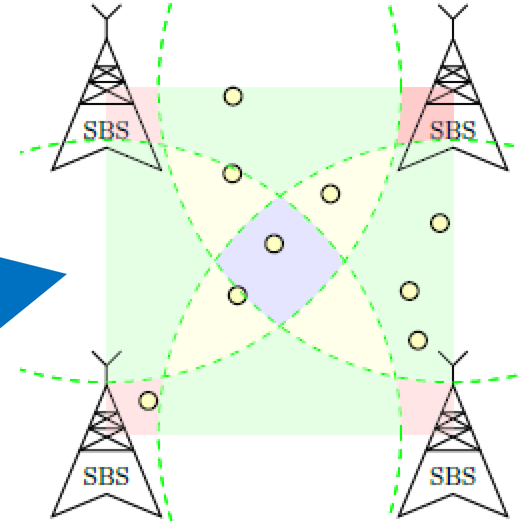


- Broadband connection is finite

Caching

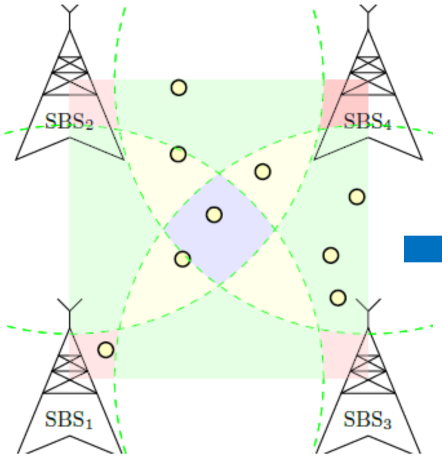
Distributed Caching in HetNets

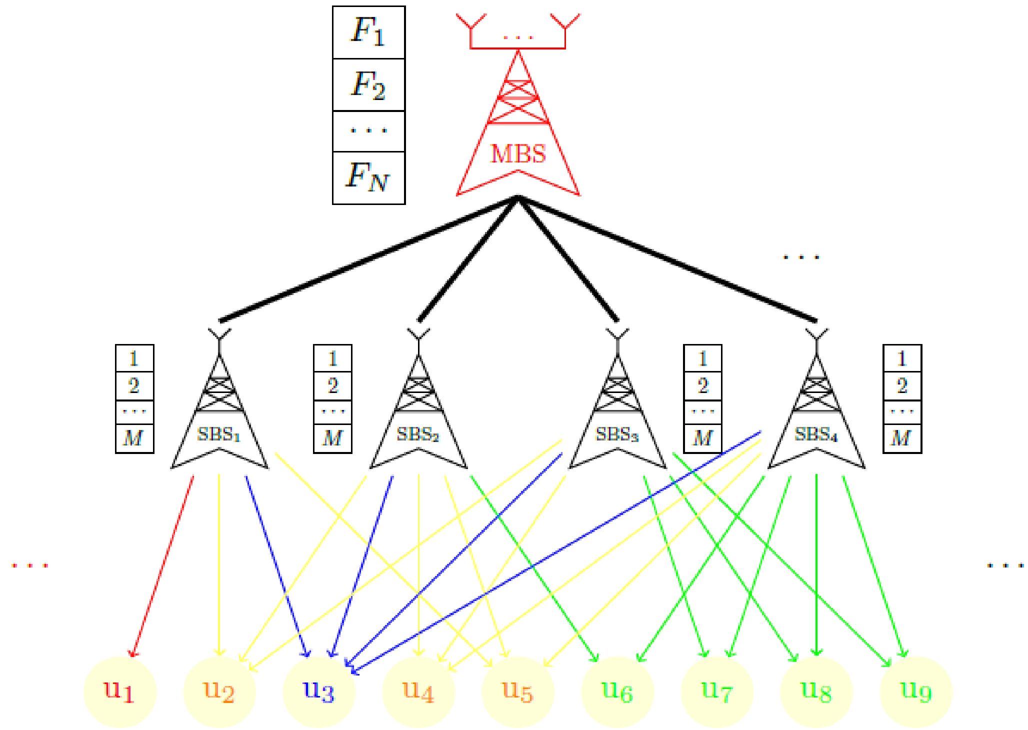




- Heterogeneous Networks:
  one **macro-cell base station** and K **small-cell base stations**
- Mobile users request files
- MBS holds N files and
  each SBS can **cache** M < N files

Distributed Caching: Graph representation





Problem: How do we fill the caches with files (or coded fragments of files) such that backhaul transmissions (MBS->SBS) are minimized?

Coded storage and delivery in HetNets

Code structure – algebraic properties

- Data file is partitioned into K data fragments
- K data fragments are mapped (encoded) to N coded fragments, N>K (vector spaces over finite fields)
- Encoding such that
  1. Original data can be reconstructed from any K coded fragments (minimum distance separable codes), or such that
  2. Original data can be reconstructed from sufficiently many K’ coded fragments (rate-less codes, fountain codes)
- Code have to support specific graph topology

Data placement in caches (SBS) – stochastic optimisation

- Data requests by mobile users to be served mainly by SBS (caches)
- Random geometric positions of users within macro cell
- Stochastic distribution of file popularity

1. Jointly profiling locations/users & contents

- Cell population small ⇒ local trafic very different from global trafic

  ⇒ requests

  behaviors/interests/profiles
- However, hard to learn with small
- Idea: jointly profile locations/users

  recommender system)

$$A = \left(\begin{array}{cc|c|c} r_{l_1,c_1} & r_{l_1,c_2} & \cdots & \cdots \\ r_{l_2,c_1} & r_{l_2,c_2} & & \\ \hline \vdots & & \ddots & \\ \hline \vdots & & & \ddots \end{array}\right)$$

Similar profiles

- $r_{l,c}$ is the history of requests at location *l* for content *c*
- « Adjacency » matrix of histories of requests:

2. Predicting Popularity via Dictionary Learning

- **Training Data** $x_{k,i} \in \mathbb{R}^m, \; i = 1, \ldots N, \; k = 1, \ldots, K$
- **Dictionary Learning**

$$\min_{A \in \mathbb{R}^{m \times L}, w \in \mathbb{R}^L} \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n} \frac{1}{2} \|x_{k,i} - A w_k\|^2 + \lambda \|w_k\|_1$$

•Training data correspond to past videos
•The entries of the feature vector represent the time evolution and metadata information

Can be solved in an online & distributed fashion

- **Vector with unobserved entries**

$$\min_{w} \frac{1}{2} \|M(\tilde{x}_k - \hat{A} w_k)\|^2 + \lambda \|w_k\|_1$$

Mask Matrix

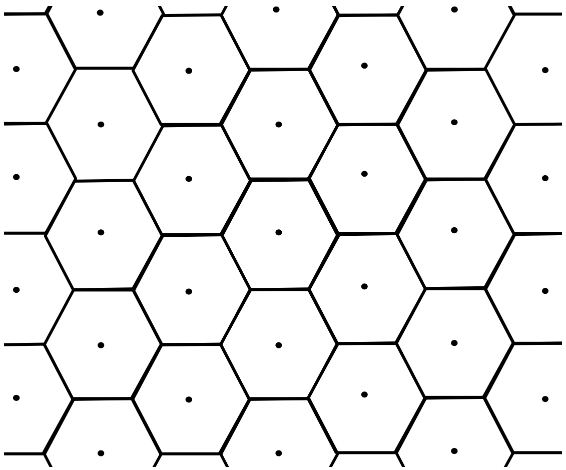
Fixed Dictionary computed previously

- **Prediction:** $\hat{x}_k = \hat{A} \hat{w}$

Questions:
•What kind of feature vectors should we use?
•Which other techniques could we use for predicting the popularity?

Structured Lattice Codes for Distributed Compression

- **Problem setup**
  - Base stations observe highly correlated signals, and are connected to central unit via finite capacity links
  - Need efficient distributed compression mechanisms



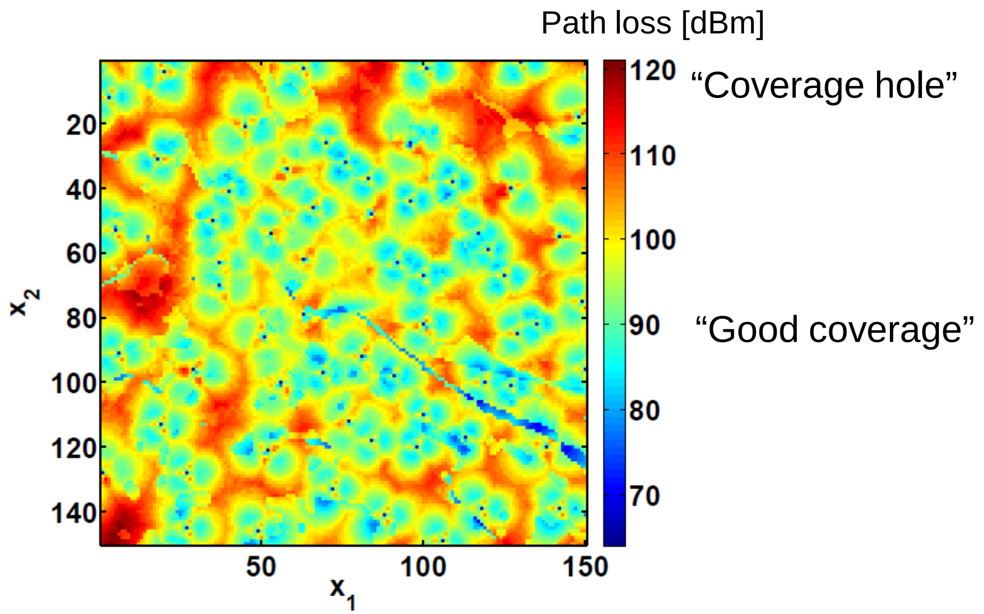
- **Goal: design feasible structured codes that approach closely the fundamental limits**
  - Structured codes (e.g., nested lattices) have the desired binning features
  - Structured codes can outperform the best known random codes in certain cases (e.g., Korner-Marton modulo-sum problem)

Page 51

Anticipatory Networks

Technical problem

- Cellular networks are full of “coverage holes”
- In these areas the service’s rate requirement cannot be supplied by the radio link



Coverage map for downtown Berlin, Germany

- If a mobile user performs media streaming (audio, video) in these areas, the channel rate is usually insufficient to support high streaming quality
- After some time, the user’s local buffer runs empty and the media stream freezes

Idea

- Assume that we can predict that a user will move into a coverage hole, then:
  - We can allocate a large portion of the wireless channel resources (e.g., bandwidth) to the user before the coverage hole
  - The user then uses this "fast" link to fill its local playback buffer
  - Within the coverage hole, the user plays the media stream from the local buffer

- Two benefits:
  1. Users receive a fluent media stream even without coverage
  2. Users within coverage holes do not need to be served by the network: Resources can be allocated to other users instead, which increases spectral efficiency of the network
- Two problems:
  1. How to predict a coverage hole, i.e., a low state of the wireless channel?
  2. How do we perform optimal resource allocation based on such prediction?

Optimal resource allocation based on prediction

- Several variants of such anticipatory resource allocation have been studied
- One of the simpler ones [1]: Linear program to adapt the fraction of resources $w_{k,t}$ allocated to user k at discrete time t

$$\min_{\mathbf{x_k}} \sum_k \sum_{t=1}^{T} w_{k,t}$$

subject to:

$$\begin{array}{lll} \text{C1:} & \mathbf{A}\mathbf{x}_k - \mathbf{V}_k = \mathbf{0} & \forall k, \\ \text{C2:} & 0 \leq \sum_k w_{k,t} \leq N & t = 1, 2, 3, \ldots T, \\ \text{C3:} & 0 \leq z_{k,t} \leq Z_k & t = 2, 3, \ldots T \ \ \forall k, \\ \text{C4:} & r_{k,t} = w_{k,t} T_d B \log_2 \left(1 + \frac{P|\hat{h}_{k,t}|^2}{N(\sigma^2 + I)}\right), & \end{array}$$

- Where:
  - Constraint 1: Assures a full buffer according to previously defined linear model
  - Constraint 2: Imposes feasibility by not allocating more than N PRBs
  - Constraint 3: Assures that the maximum buffer size is not exceeded
  - Constraint 4: Updates the rate values in C1, only given for clarity

[1] S. Sadr and S. Valentin, "Anticipatory Buffer Control and Resource Allocation for Wireless Video Streaming", *arXiv:1304.3056 [cs.MM]*, Apr. 2013.

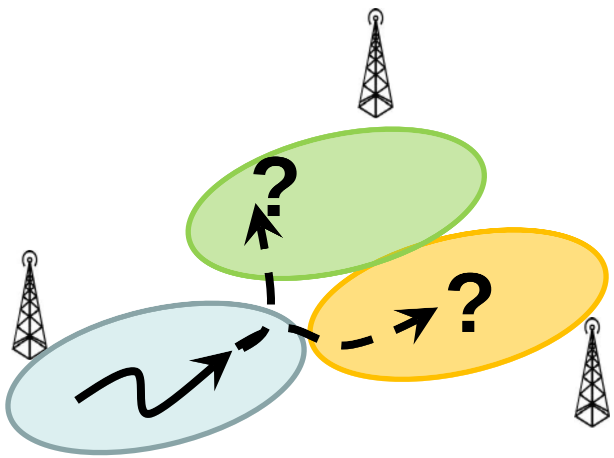
Required research

- Problems of current formulations:
  1. They do not account for prediction error: An error norm should be integrated into the optimization problem
  2. They do not schedule in time: Instead of current instantaneous optimization, resources should be allocated along the time axis
- Note: Strict real-time constraints apply, 100s of users must be served within 1 ms

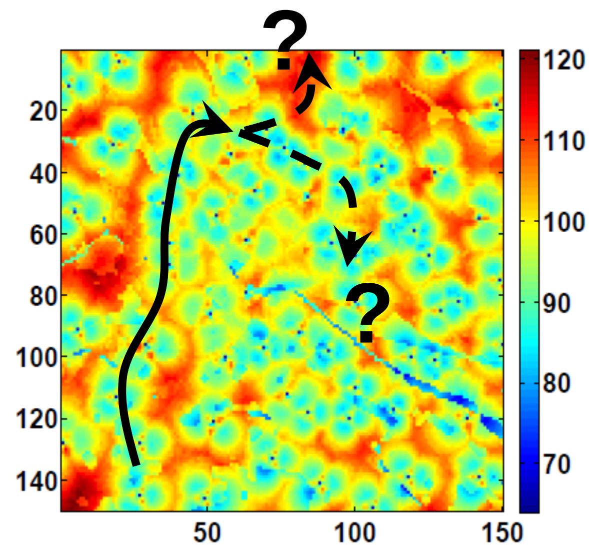
- Research tasks:
  1. Find norms for the prediction error that maintain stable and fast optimization, if they are included into the optimization problem
  2. Extend current formulations by scheduling in time, while still providing real-time solutions for multiple users
- Hint: Exploiting the high spatio-temporal correlation of the wireless channel may help here, e.g., to apply parametric programming

Mobility prediction

- Mobility prediction can be used as a primitive for many optimizations in cellular networks



Handover optimization

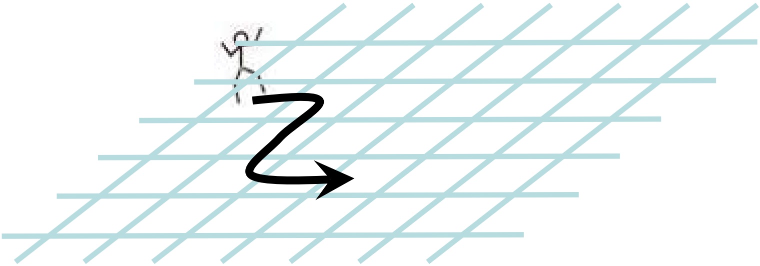


Signal & QoS prediction and anticipation

- However, typically difficult to gather enough individual data to make accurate prediction (due to privacy, battery usage, lots of data…)
- Our approach is to leverage other users' data to predict the mobility of a given user

Collaborative mobility prediction

- **An example of method: hierarchical Bayesian inference using DP mixture priors:**
- **Quantize space according to application needs (based on cell coverage, radio maps, etc)**



- **Define:**
  - Set of feasible patterns, ex: transition to adjacent cells
  - Mobility model, ex: 1-st order Markov: $\mathbb{P}\left(u \text{ moves from } i \text{ to } j\right) = \theta^u_{i,j}$
  - High-level prior on mobility kernels, ex: $g = \sum_k \omega_k \delta_{\theta_k} \sim \mathrm{DP}(\alpha, G_0)$

  (any sample from *g* defines a clustering of users with similar mobility)
- **Perform Bayesian updates on the cluster structure**

  ⇒ **aggregation brings statistical power for learning mobility**

  ⇒ **Bayesian framework allows adaptivity of number of clusters and soft-clustering**

Radiomap Reconstruction

- Coverage map of a certain area represented as an matrix $\boldsymbol{H} \in \mathbb{R}^{N \times N}$
- We have at our disposal a small subset of the original n denoted by

$$\boldsymbol{P} \in \mathbb{R}^{N \times N}$$

- Resort to matrix completion techniques to find the missing entries. Solve the optimization

$$\min_{\boldsymbol{A}^{N \times N}} \|\boldsymbol{A}\|_*$$

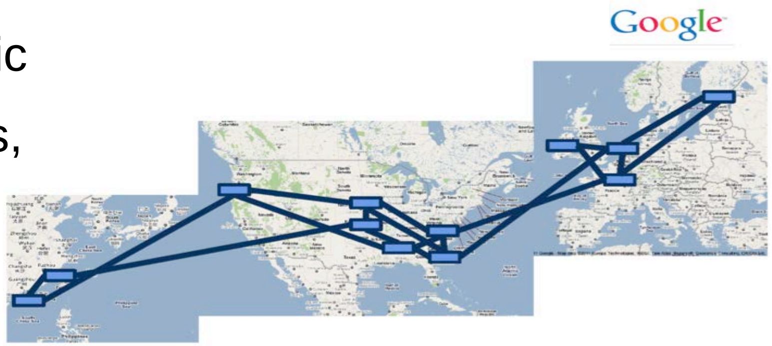
$$\text{s.t.} \quad A_{ij} = P_{ij}$$

- Many interesting applications stem from the radiomap reconstruction, e.g., accurate channel prediction exploiting trajectories, include side information (importance of a region) to improve the results, etc.

Bandwidth Calendaring

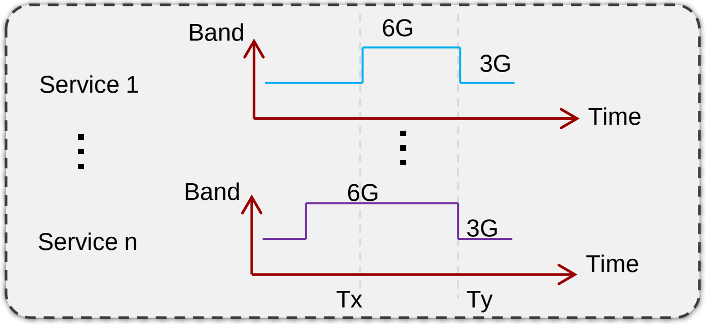
Bandwidth calendaring

- **Context: Inter-datacenter networks**
  - Deployed by cloud companies operating geo-distributed datacenters
  - Need to support bulky and predictable traffic across datacenters (map reduce operations, database synchronization, etc.)



- **Main problem**
  - Find feasible transfers in time and space such that their execution delay is minimized.
  - In an online manner (with deterministic and probabilistic knowledge of the future)
  - At scale (high demands, large networks)

Bandwidth Calendar



- **State of art**
  - Earliest Deadline First (EDF) far from optimal
  - Offline impractical (scalability, future knowledge)

Online deterministic problem formulation

For each Request i

$$\begin{cases} s_i : \text{source} & t_i : \text{destination} \\ b_i : \text{start time} & P_i : \text{admissible paths} \\ d_i : \text{deadline} & \\ D_i : \text{demand size} & f_{i,p,t} : \text{request i on path p at time t (variable)} \end{cases}$$

$$\max \sum \sum \sum U_{i,p,t} . f_{i,p,t} \quad \text{maximize utility at every step } \tau$$

while maximizing the minimum execution delay (α_i)

Two stage optimization

$$st. \sum_{i;\max(\tau,b_i)\le t\le d_i} \sum_{e\in p} f_{i,p,t} \le c_e \ \forall e \text{ (edge capacity)}$$

$$\sum_{i;\max(\tau,b_i)\le t\le d_i} \sum_{p} f_{i,p,t} \le D_i - F_{i,\tau-1} \quad \forall i \text{ (remaining demand)}$$

amount of demand i satisfied at previous step

$$\sum_{i;\max(\tau,b_i)\le t\le d_i}^{d_i} \sum_{p} f_{i,p,t} \ge \alpha_i D_i - F_{i,\tau-1} \quad \forall i \text{ (execution timeliness)}$$

maximum delay tolerance of demand i

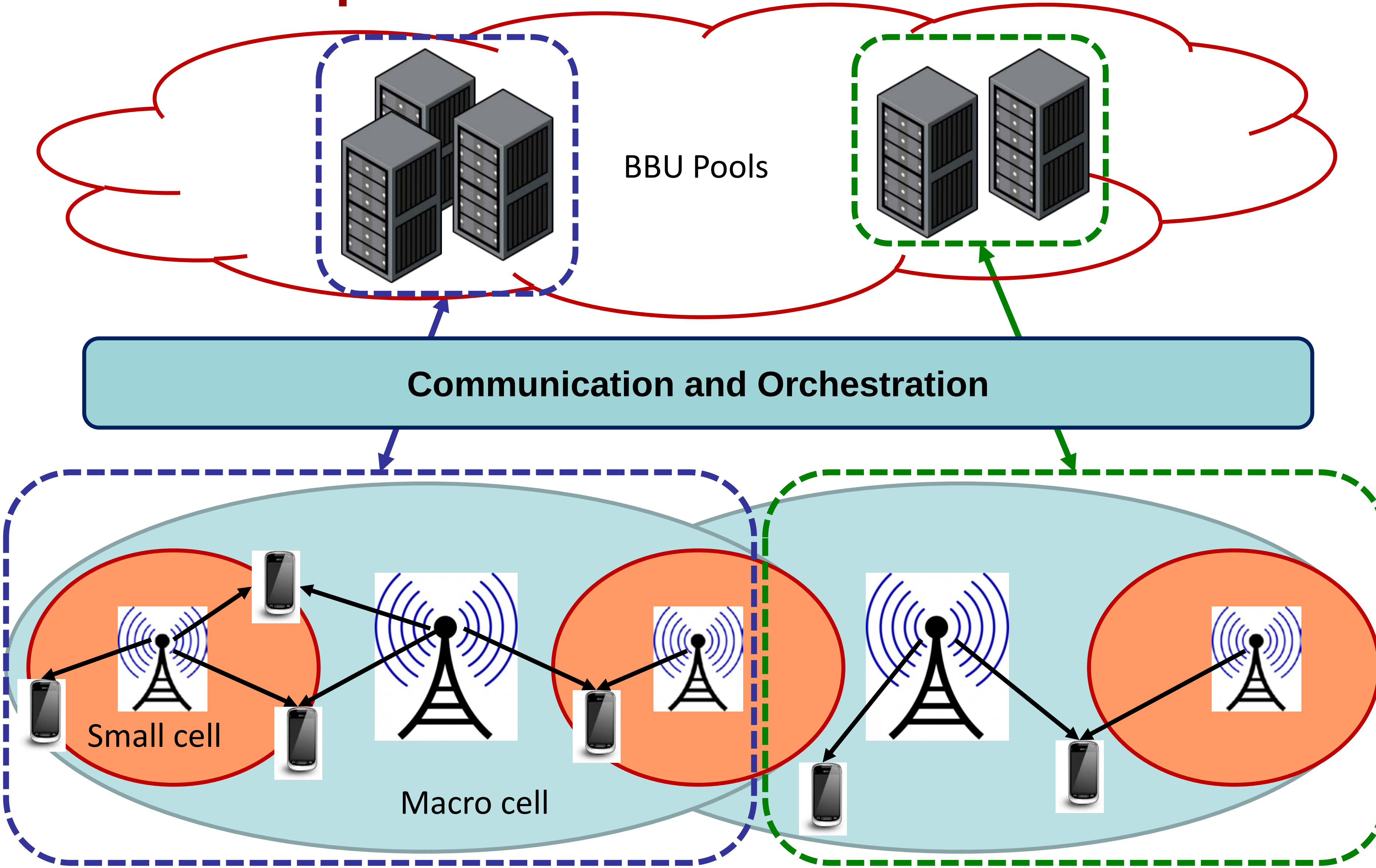
$$\sum_{i} \sum_{\max(\tau,b_i)\le t\le d_i}^{d_i} \sum_{p} U_{i,p,t} . f_{i,p,t} \ge U - U_{\tau-1} \quad \text{(min utility)}$$

Reference: S. Kandula et al. Calendaring for Wide Area Networks. Sigcomm 2014.

Online optimization dealing with uncertainty

- **System considerations**
  - Consider **stochastic** inputs (e.g., demands)
  - Ensure fairness between demands (max min α for instance)
  - Leave maximum space for future demands
  - Use iterative solving methods (online systems/algorithms)
  - Handle millions of variables and constraints
- **Training objective**
  - Formulate bandwidth calendaring problem taking into account the uncertainty (e.g., stochastic demands, failures, etc.)
  - Propose an iterative method for solving large problem instances
    - Possibly based on *stochastic programming* and/or *robust optimization*

Case 3: Spectrum virtualization for C-RAN



Research items on C-RAN

- **Radio Resource Coordination**
  - Dynamically allocation of contiguous and non contiguous bands for dense deployments
  - Joint scheduling of transmissions and assignment of radio resources
  - Maximizing the overall throughput, reducing channel switching, reducing energy-consumption, reducing interference
- **BBU pool management**
  - Dynamic allocation of BBU processing to minimize energy consumption and maximize throughput on intra and inter-BBU pools
  - Considering specific constraints of backhauling (cost, characteristics, congestion)
- **System consideration:**
  - Dynamic UE-driven RRH clustering (user mobility)
  - Exploit knowledge and learn from the past (temporal and spatial variation of traffic)
  - Consider stochastic inputs (be robust against small input variation)
  - Exploit network diversity (Coordinated Multi-Points)

Radio Resource Coordination

- **Main problem**
  - Design mechanisms to dynamically allocate contiguous and non contiguous bands depending on the load for dense deployments
  - Maximize the overall throughput, reduce of channel switching, reduce energy-consumption

- **Methodology**
  - *Decision* and *control theory* (MDP) to model the system.
    Regard spectrum as *bandits* (possibly correlated) and incorporating the specific constraints on observation errors, bands, energy efficiency.
  - Propose *approximation algorithms* (PTAS) with possibly bounded performance gaps.
  - Define *provably efficient protocols* to manage spectrum handovers.

BBU pool management

- **Main problem**
  - BBU processing has limitations in number of bands and number of RRHs
  - Dynamically allocate BBU processing to minimize energy consumption on intra and inter-BBU pools
  - Integrate specific constraints on backhauling (cost, characteristics, congestion) for inter-pool case.

- **Methodology**
  - Define and solve *optimization* models (MILP) using *stochastic* programming and *robust* optimization to consider uncertainty (intra-pool)
  - Propose *scalable heuristics* for radio resource allocation and transmission scheduling
  - Partially decentralized schemes using *game theory* (inter-pool, decentralized case)

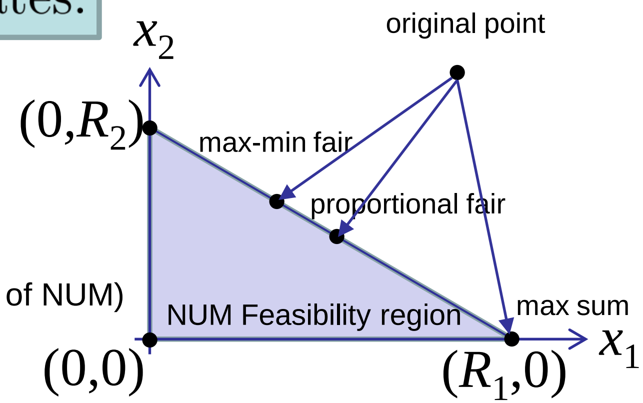
Network Control Optimization

Utility Maximization and Fairness in Base Stations

- **Modern base stations utilize convex optimization algorithms to share resources**
  - Users are interested in communication rates, x_i denotes the rate of user i in Mbps
  - We model the user satisfaction with a concave utility function $U_i(x_i)$
    - The concavity of U_i is characteristic of diminishing returns "*a well-served user gains little by adding more rate*"
  - The base station strives to optimize the sum of utilities

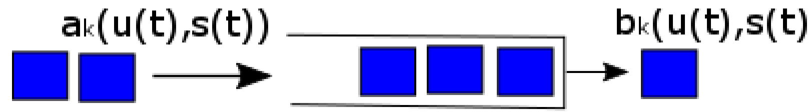
NUM Problem: $\max_{(x_i)} \sum_i U_i(x_i)$ s.t. x_i are feasible rates.

- **Fairness**
  - Vectors (x_i) are fair if they satisfy specific criteria that reflect fairness
    - **Max-min fair**: cannot improve poorest user without making another user even poorer
      - Always exists and is unique in convex sets (like the feas. region of NUM)
    - **Proportionally fair**: users receive rate based on their capabilities
      - Improves utilization of network resources
    - **Max sum rate**: maximum utilization but possible starving some users



- **Connection between convex optimization and fairness**
  - Family of alpha-fair functions $U_i(x) = \frac{x^{1-\alpha}}{1-\alpha},\ \alpha > 0$
  - Different choices of *α* lead to different functions. The solution of the corresponding NUM optimization yields the desirable operational points. Max sum rate (α=0), Proportional fairness ($\alpha \to 1$), Max-min fairness ($\alpha \to \infty$).
- **Stochastic NUM captures the problem at a packet level**
  - x_i are time averages of transmissions, the original input point is adaptively driven to the desired fairness point (see fig)

Stochastic Network Optimization: General formulation



a data queue – arrivals and service depend on control $\mathbf{u}$ and state $\mathbf{s}$

- **The stochastic optimization problem**
  - Consider a network with data queues:

    $$Q_k(t+1) = [Q_k(t) - b_k(\mathbf{u}(t),\mathbf{s}(t))]^+ + a_k(\mathbf{u}(t),\mathbf{s}(t))$$

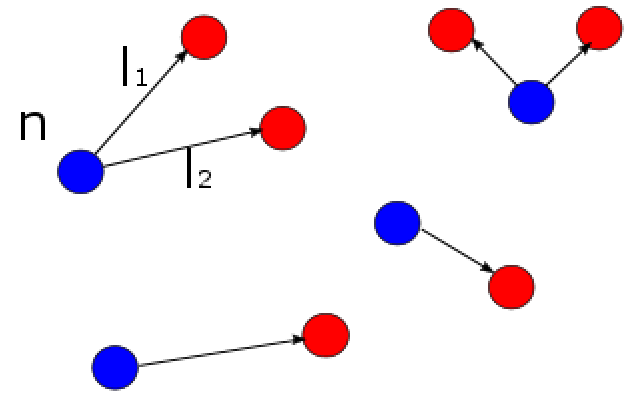
  - **u**(t): control at t (e.g. schedule, forwarding decisions, power control etc.)
  - **s**(t): uncontrolled system states at t (e.g. channel conditions, link capacities etc.)
  - **x**(t) = [x_1(**u**(t), **s**(t)), …, x_M(**u**(t), **s**(t))], **y**(t) = [y_1(**u**(t), **s**(t)),…, y_L(**u**(t), **s**(t))] : Attributes associated with control decision and the state of the system, e.g. power cost, reward for admitting a packet etc.
    - Define the time average of quantity z(t) as $\bar{z} = \limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}(z(t))$
  - **Stochastic Problem:** Find a dynamic control policy **u**(t) that solves:
    - *Maximize* $f(\bar{x})$ *(f is a convex function)*
    - s.t. $\bar{y}_l \le 0, \forall l = 1,\dots,L$

      all data queues are mean rate stable, i.e. $\lim_{t\to\infty} \frac{\boldsymbol{E}(Q_k(t))}{t} = 0$

- **Lyapunov Optimization:**
  - Auxiliary variables **v**(t)=[v_1(t),…,v_M(t)],control parameter V, virtual queues Z_l(t+1)=[Z_l(t)+v_l(**u**(t),**s**(t))]$^+$, G_m(t+1)=[G_m(t)+v_m(t)-x_m(**u**(t), **s**(t))]$^+$, Lyapunov function $L(t) = \frac{1}{2}\left(\sum_{t=1}^{K} Q_k(t) + \sum_{m=1}^{M} G_m(t) + \sum_{l=1}^{L} Z_l(t)\right)$
  - At each slot, observe **Q**(t), **G**(t), **Z**(t), **s**(t) and choose **c**(t), **u**(t) to minimize (an upper bound of) the « Drift plus Penalty » $\boldsymbol{E}\left(L(t+1) - L(t) \middle| \mathbf{Q}(t),\mathbf{G}(t),\mathbf{Z}(t)\right) - V\cdot\boldsymbol{E}\left(f(\mathbf{c}(t)) \middle| \mathbf{Q}(t),\mathbf{G}(t),\mathbf{Z}(t)\right)$

Stochastic Network Optimization: Application

- **Application: Energy Optimal Control for Time Varying Wireless Networks [M.J. Neely, IEEE Trans. IT, 2007]**
  - **System Model:** N nodes, L links, L_n is the set of links where n is a source, $Q_l(t)$ is the queue for data to be transmitted at link l.
    - State of the system **s**(t) : The channel states at slot t, **H**(t) and the exogenous arrivals for link l, $A_l(t)$.
    - Controls: The transmission power $P_l(t)$ for each link l, admission control at the nodes ($a_l(t)$ the data admitted at $Q_l(t)$).
    - Rate at link l: $b_l(\mathbf{P}(t),\mathbf{H}(t))$,
    - average power used by node n: $\bar{p}_n = \limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}\sum_{l\in L_n} \boldsymbol{E}\left(P_l(t)\right)$
    - average amount of data for link admitted at node n $\bar{a}_l = \limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}\sum_{l} \boldsymbol{E}\left(a_l(t)\right)$



Use wireless transmission scheduling with power control to minimize the average consumed energy for given throughput

- **Problem: Maximize weighted sum throughput $\sum_l \theta_l \cdot \bar{a}_l$ subject to power consumption constraints $\bar{p}_n \le P_{n,av}$ and stability of all $Q_l(t)$.**
- **Algorithm: Define virtual queues for the power constraints:** $Z_n(t+1) = \left[Z_n(t) - P_{n,av} + \sum_{l\in L_n} P_l(t) \right]^+$
  - Main idea: If the virtual queues are kept stable, power constraints are met.

1. *Admission Control phase* : If $Q_l(t)<2\theta_l/V$ admit all arriving data $A_l(t)$, drop them otherwise.
2. *Power Control phase* : $\mathbf{P}(t) = \arg\max_{\mathbf{p}\in\Pi} \sum_{n=1}^{N}\sum_{l\in L_n} \left(Q_l(t)\cdot b_l(\mathbf{p},\mathbf{C}(t)) - Z_n(t)\cdot p_l \right)$

Main result: The constraints are satisfied while the utilityobtained is near optimal

Thank you

www.huawei.com