

# From Embedded Systems to Scalable Platforms: Challenges in the Development of 5G baseband system on chip

Alan Gatherer PhD Senior Technical Vice President, Huawei USA

[alan.gatherer@huawei.com](mailto:alan.gatherer@huawei.com)

[www.huawei.com](http://www.huawei.com)

HUAWEI TECHNOLOGIES CO., LTD.



# Huawei: a brief summary

**170+**  
Countries



**14**  
R&D Centers

**31**  
Joint Innovation Centers

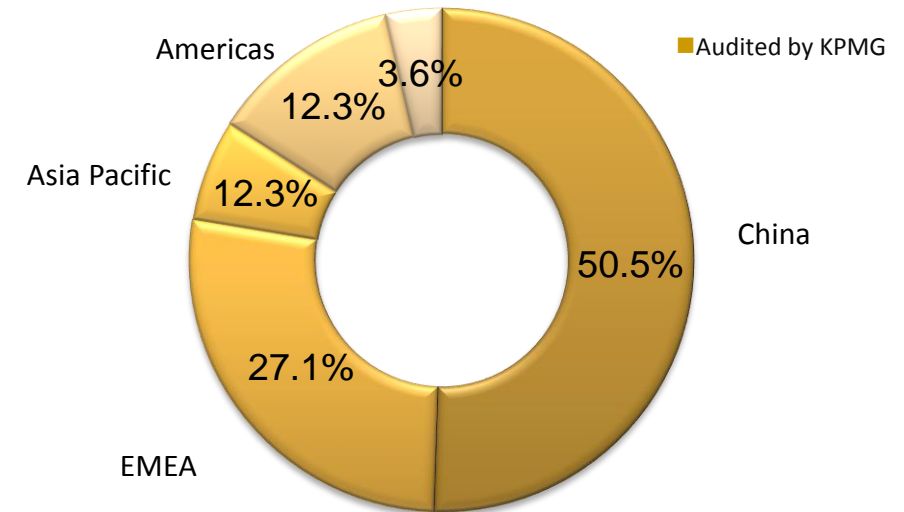
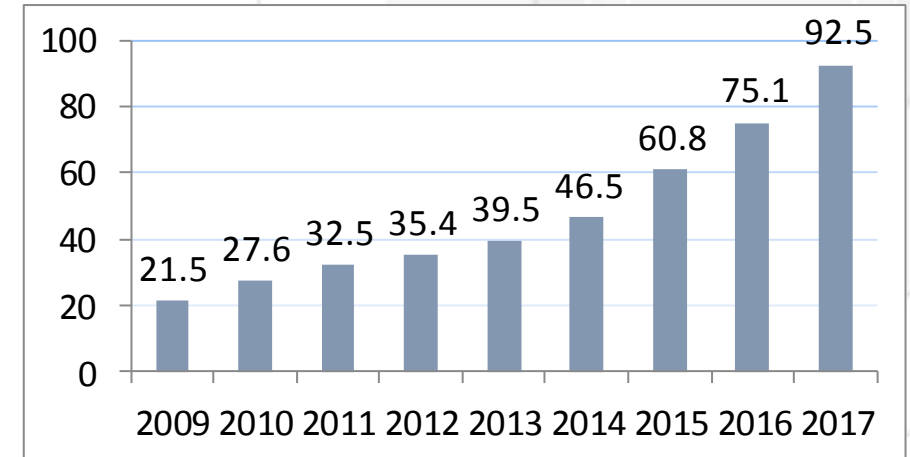
**14**  
Regional HQs



**45**  
Training Centers

**180,000**  
Employees Worldwide

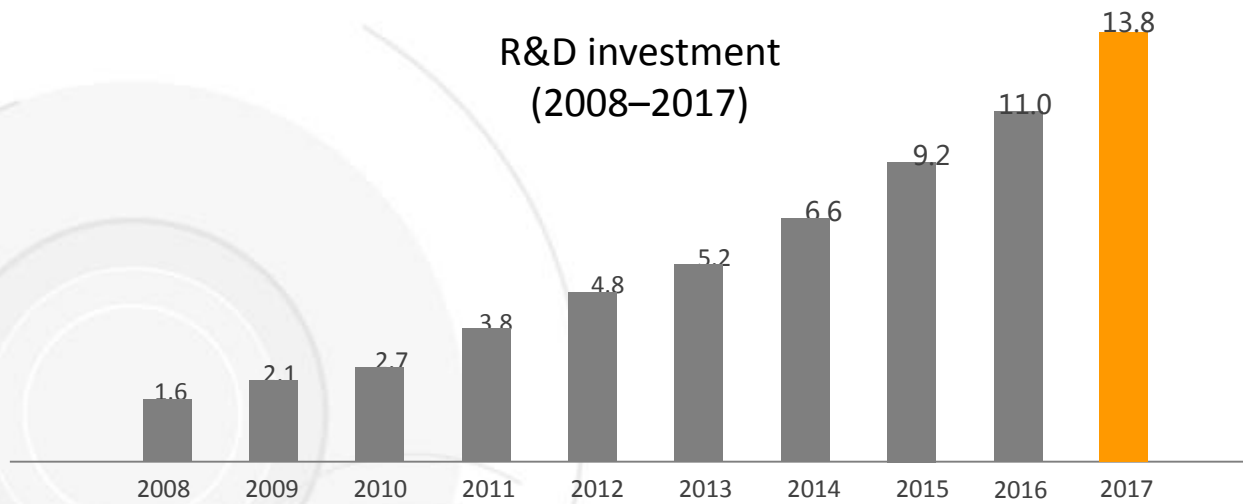
Revenue in \$B USD



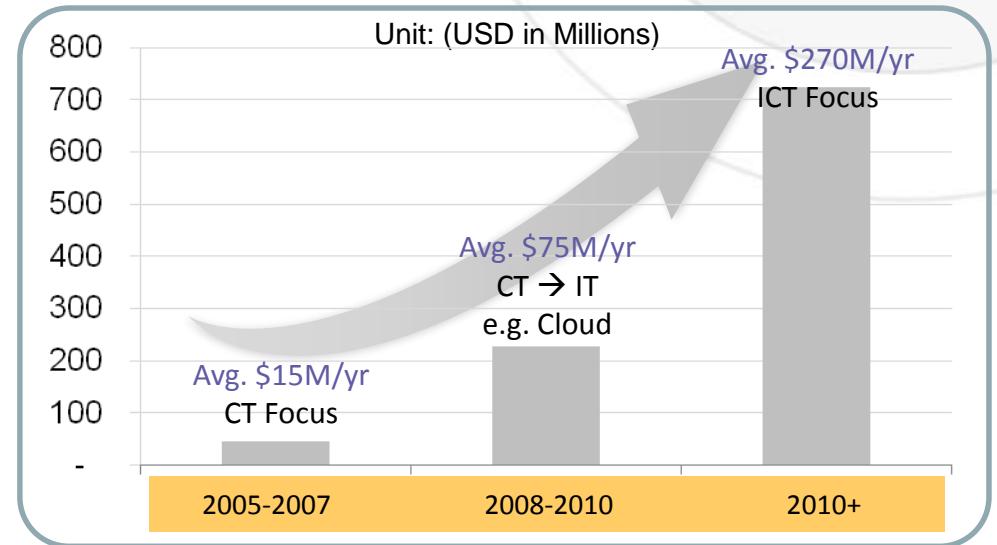
# Research Investment

Increase in Strategic Investments and Customer Focus Innovation

Continuous Innovation Investment



R&D Investment in USA



# U.S. R&D Offices

- Established the first US R&D office in 2001,
- HQ-ed in Santa Clara, CA
- ~900 staff in 9 R&D offices across USA



Office



University Connections



US R&D HQ  
Santa Clara



## WEST

- UC Berkeley
- Stanford U
- UCSD
- UCLA
- UC Irvine
- UC Davis
- UC Riverside
- CalTech
- Santa Clara U
- UCSB
- UC Santa Cruz
- USC
- OSU
- WSU
- U of Washington

## SOUTHWEST

- Texas A&M
- UT Dallas
- UT Austin
- U of Arizona
- Arizona State U
- U of Houston
- Rice University
- Univ. North Texas
- SMU
- UT San Antonio
- UTEP
- U of Colorado

## MIDWEST

- Purdue U
- Univ. Minnesota
- Washington U
- U of Michigan
- U of Missouri
- U of Wisconsin
- UIUC
- U of Notre Dame
- Northwestern U
- Wayne State U
- Ohio State U
- Drexel U
- IIT
- U of Chicago

## SOUTHEAST

- NCSU
- Georgia Tech
- University of Florida
- Duke University
- UNC-Chapel Hill
- U of Virginia
- Virginia Tech
- Clemson U
- FIU

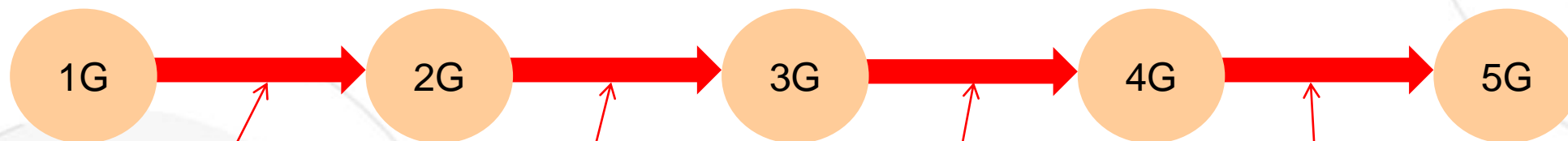
## NORTHEAST

- SUNY – Bingham
- SUNY Buffalo
- U of Delaware
- U of Maryland
- Rutgers
- U of Rochester
- Manhattan College
- Tufts U
- NYU
- Princeton
- U of Penn.
- MIT
- CMU
- Cornell U
- Harvard U
- Columbia U
- Yale U

# Part I: 5G the basics

# 5G: A Unique Time for Cellular Standards!!

Why?? We all actually want this one!!

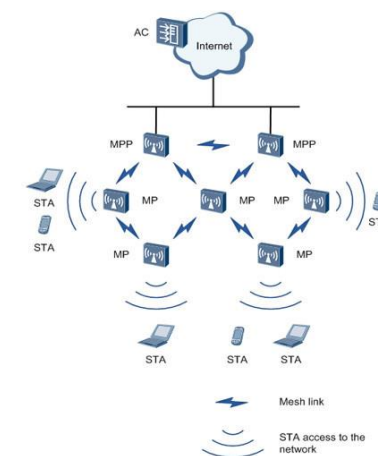


2G, cost reduction, really

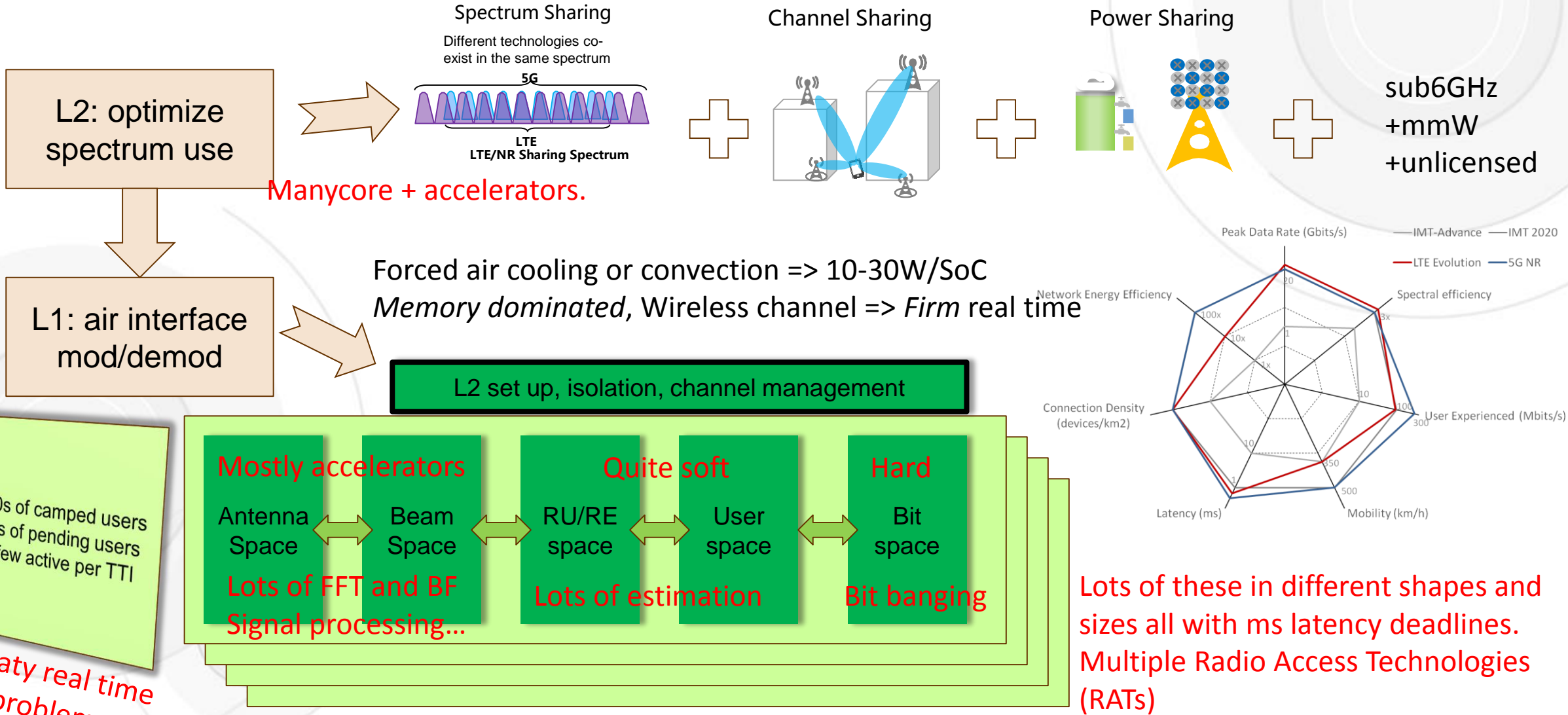
3G, data, but why do we need it?  
Voice is good!!

4G, fix 3G.  
Ooops..

5G. Yes I want that!!  
..lots of new market opportunities in V2V, IoT etc. Maybe we finally understand data? AND we have competition in a way we have never had it before



# 5G baseband in a nutshell



# Emerging Market and Technology Drivers



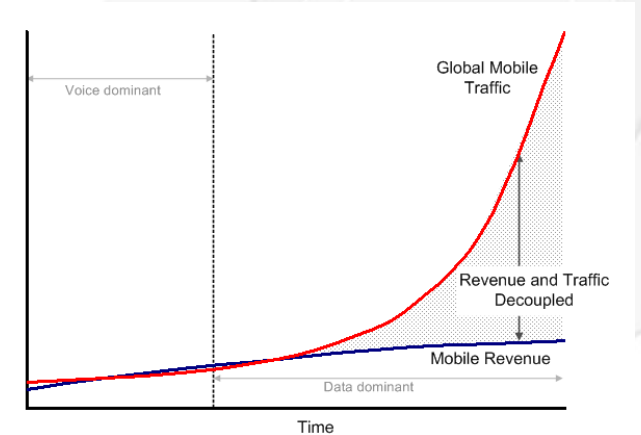
*RAT = Radio Access Technology*

*RAN = Radio Access Node*

## Multi RAT, RAN as a Service

- A good schedule for one RAT may not make for a good schedule when another RAT is added. So we have a basic **RAT mixing problem** that has the potential to lose us significant performance if done badly (either in Normalistan or Extremeistan)

Need to start playing the statistics of RAT use to save costs



Traffic data from "Recognizing the promise of mobile broadband", UMTS Forum, 2010  
Revenue data from "Global mobile broadband: market potential for 3G LTE", Analysys Research Limited, 2008

## Requirement uncertainty from 5G

- leading to a softer modem and the need find strategies to mitigate the cost of softer IP. **Scheduling** becomes a key technology

## New network topologies

- CRAN, MEC, Enterprise, Neutral Hosting must all be supported using a **single software architecture** strategy

## Software Reuse

- General cost reduction in maintenance of baseband across multiple platforms is an issue that becomes more significant as we move to 5G and mRAT

## End of Moores Law

- Performance improvement becomes unreliable. We need to find other paths to system performance improvement



# Baseband: Not an Embedded System, not a Compute Server

Normalistan

Extremeistan

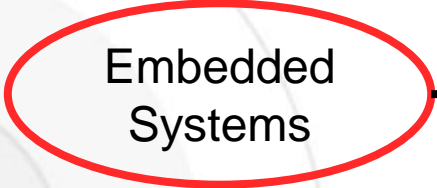
- Traditionally we have treated it like an embedded system
  - Embedded system level of flexibility in architecture.
  - Software upgrade mainly to fix bugs

- Compute Servers will be used in BBSoC architecture
  - close to baseband in architecture requirement?
  - addition of ML may make them more “computish”
  - the industry effort is massive!



Application Requirements known

Scale up



Scale out

IP Performance optimized for algorithm

Application Requirements unknown

Silicon optimized

## The road from Normalistan to Extremeistan

- What issues prevent its immediate application?
- What added value do we get from it?

**SUPPORT OF HETEROGENEOUS IP AND MEMORY  
FIRM REAL TIME REQUIREMENTS**

**SCALE OUT, FLEXIBILITY, COMPOSABILITY, SCALABILITY, VIRTUALIZATION  
GENERAL ALIGNMENT WITH CLOUD ARCHITECTURE. INTUITIVELY A GOOD THING..**

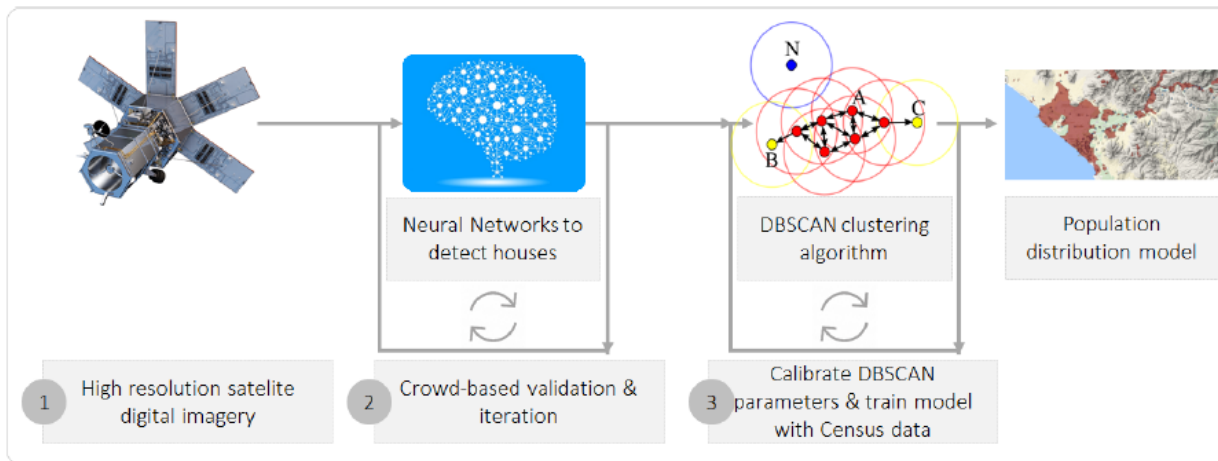
# ORAN, TIP and so on

## xRAN Forum Merges With C-RAN Alliance to Form ORAN Alliance

February 27, 2018 03:55 AM Eastern Standard Time

BARCELONA, Spain--(BUSINESS WIRE)--The xRAN Forum today announced its intent to merge with the C-RAN Alliance to form a world-wide, carrier-led effort to drive new levels of openness in the radio access network of next-generation wireless systems. The work of the ORAN Alliance will combine and extend the work of both the C-RAN Alliance and the xRAN Forum, while maintaining the key objectives of each group.

How is isolated population detected without reliable census data?



The goals of the OTTs are different from the traditional operators

- Find new customers to mine
- Increase marketing opportunities
- A data mining approach
- Not too fixated on cost of hardware



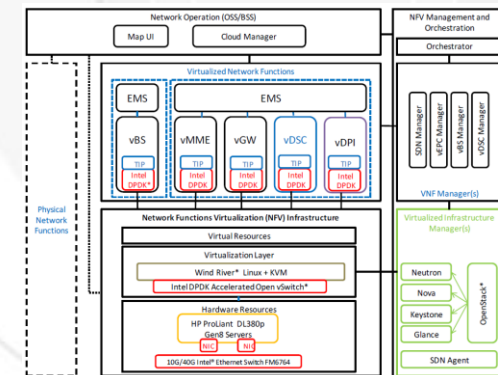
Reuters/Elijah Nouvelage

- **Autonomous cars could net telecom carriers \$1.3 trillion annually, Morgan Stanley estimates.**
- **It could be a bigger revolution than the smart phone.**
- **Still, 5G faces a unique set of challenges before it will revolutionize our commutes.**

## **Part II: Finding Patterns in a Service Oriented Network**

# Can Cellular Infrastructure be just a big cloud app?

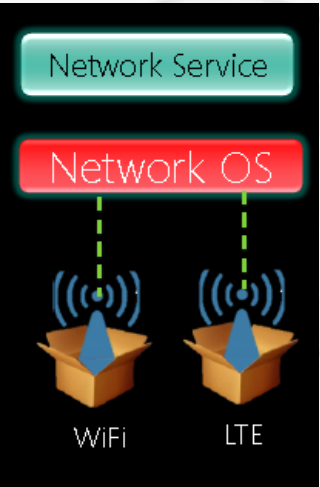
1. NFV: a classic example of virtualization, but there is difficulty in applying it to the RAT
  - How to manage real time? Nova resource is on the face of it not so efficient
  - Even Docker for RAT seems like overkill
  - Beginning to look like SDR!!
2. SDN for wireless: more useful sounding, app centric view of RAT. Integrate with SDN
  - How to deal with sounding, control and other common channels?
  - Benefit is to the app, so the benefit to the RAT is not so obvious (but that is OK)



### 3. Application/service driven network: customers pay for apps not bits

- Because that's where the money is...but how do we do that?
- In the future there will be a lot **more rented infrastructure** at multiple levels
- Some operators want to become software houses
- Actual infrastructure is just a burden; "shape of your money", why take the risk?
- Don't let the OTT eat all the good bits
- We cannot reach 1000X bandwidth but maybe we can reach **1000X experience!**
- IT says experiment but this also requires extreme agility.

In search of pragmatism. Although active networks articulated a vision of programmable networks, the technologies did not see widespread deployment. Perhaps one of the biggest stumbling blocks was the lack of an immediately compelling problem or a clear path to deployment. A significant lesson from the active-network research effort was that **killer applications for the data plane are hard to conceive.** The community proffered various applications that could benefit from

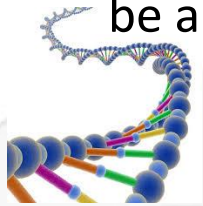


# If the Infrastructure is an organism..

**Recap:** Sell services or apps, actual infrastructure is sooo 2000s ...

**Unfortunately:** Bad infrastructure will still kill you

**Fortunately:** Good infrastructure is still a great money maker and will be an essential part of the 1000X goal



## Features are like genes

- Feature combinations may surprise us
  - MIMO for instance is not always a capacity improvement
- The environment may change the answer for a features value
- The application use may change the features value
- The environment and application space are getting exponentially more complex!!
  - Now 5G wants us to support 3 distinctly different physical requirements
- Unlike evolution, we cannot afford a blind watchmaker.....but adaptability is critical.
- **If the infrastructure is an organism then humans are the food!!** Fortunately we know something about them.



MIMO, scheduling,  
antenna tilt, IC,  
CoMP...

Features!

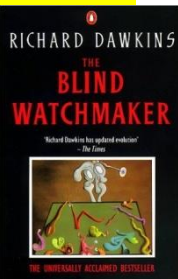
||



“can 20% CoMP increase the satisfaction index for twitter users in a suburban, network in Asia?”

“yes, but only if they use periscope 28% of the time”

“what about the next cool app?...”



# Humans: more predictable than you might think..

## Point #1: Only so many can fit in a box



- Our systems drop packets all the time it is OK to play the statistics on the hardware too.
- In fact this is one of the big arguments for CRAN

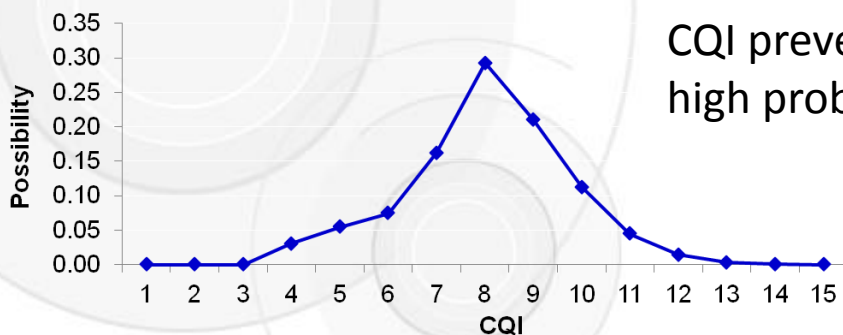
- Implication:
  - Operator sharing of resources is a guaranteed win if you are looking to reduce equipment, and yet we don't do it. This is called capitalism...
- Caveat:
  - How many IoT devices will fit in starbucks??
- Crazy idea:
  - Shared hardware resources per unit time are bought and sold in some sort of commodity market
  - Spread the risk of hardware purchase
  - Reduce redundancy, save the planet....
  - Introduce a whole new class of traders??

# Humans: more predictable than you might think..

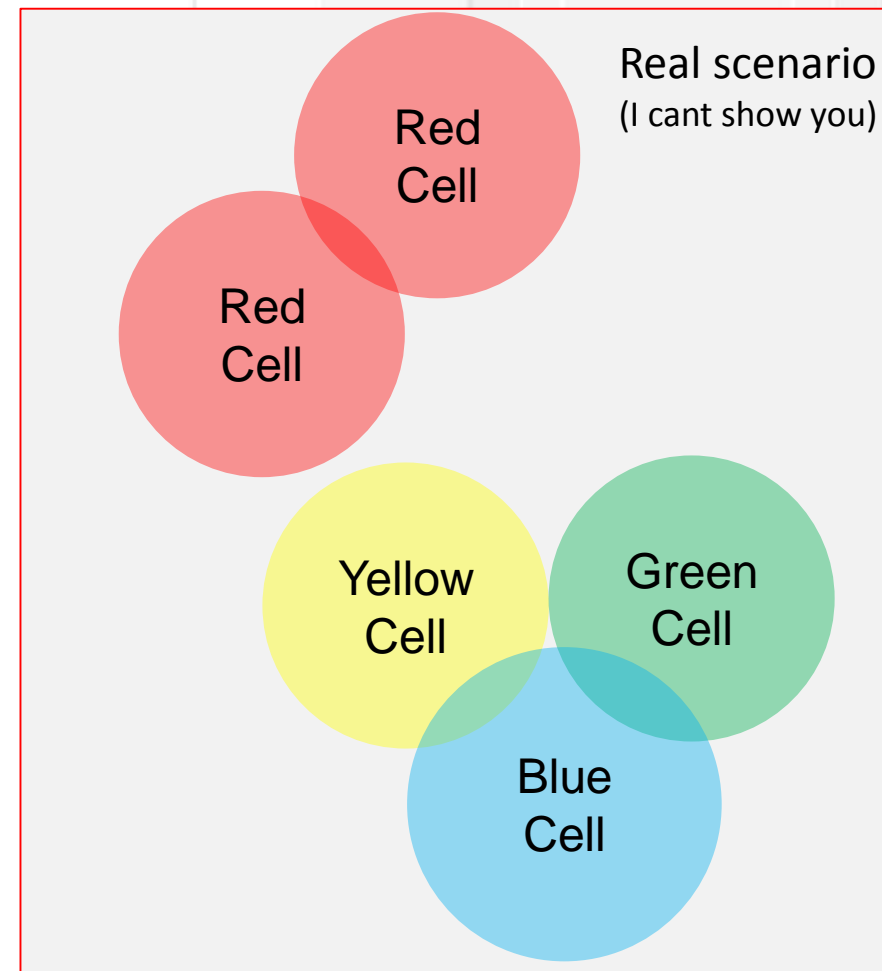
## Point #2: They don't come from nowhere

Scenario	Correlation Blue	Correlation Green	Correlation Blue+Green	Correlation Red
A	0.18	-0.56	-0.43	0.13
B	-0.25	-0.41	-0.42	-0.22

Wave effect leads to strong negative correlations over the short term between close neighbors



CQI prevents anything near "worst case" with high probability

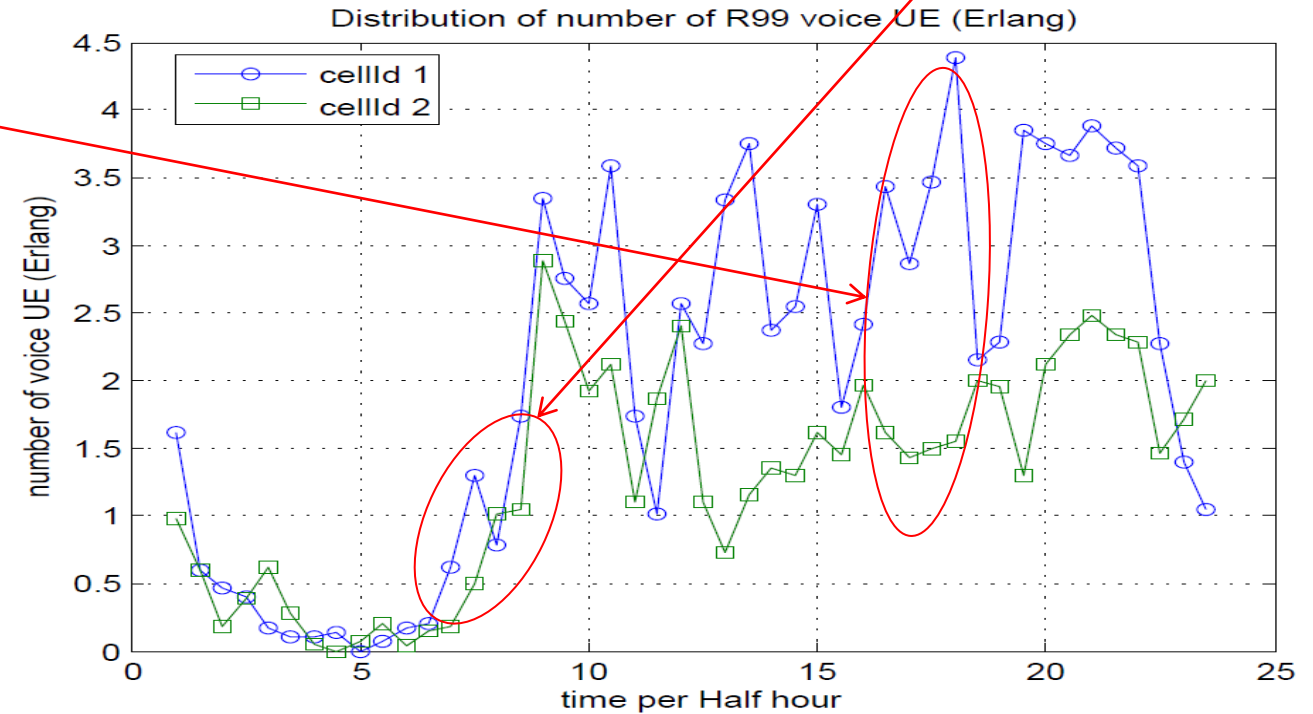


Worst case loading of network hardware depends, but 25% would not surprise and it will only reduce into 5G

# Humans: more predictable than you might think..

## Point #3: They are basically insects, or wolves...

- In real data, large area average seen due long term user migration, the **tidal effect**
- Short term **wave effect** rides on top
- **Wave effect is more useful**
  - Short distance, short term
- **Humans seem to move in well defined statistical patterns seen in bees, wolves, birds etc.**



## On the Levy-Walk Nature of Human Mobility

Injong Rhee, Member, IEEE, Minsu Shin, Student Member, IEEE, Seongik Hong, Student Member, IEEE, Kyunghan Lee, Associate Member, IEEE, Seong Joon Kim, Member, IEEE, and Song Chong, Member, IEEE

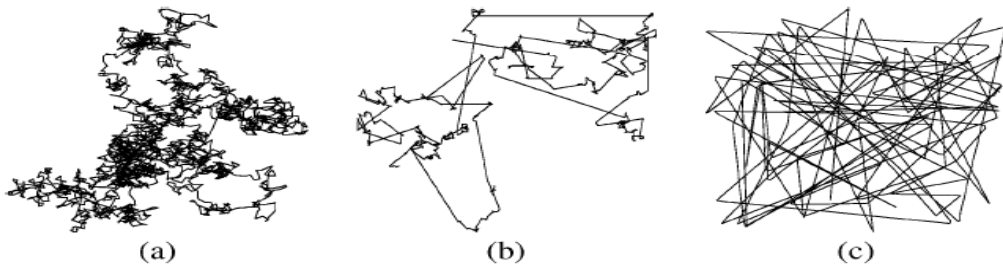
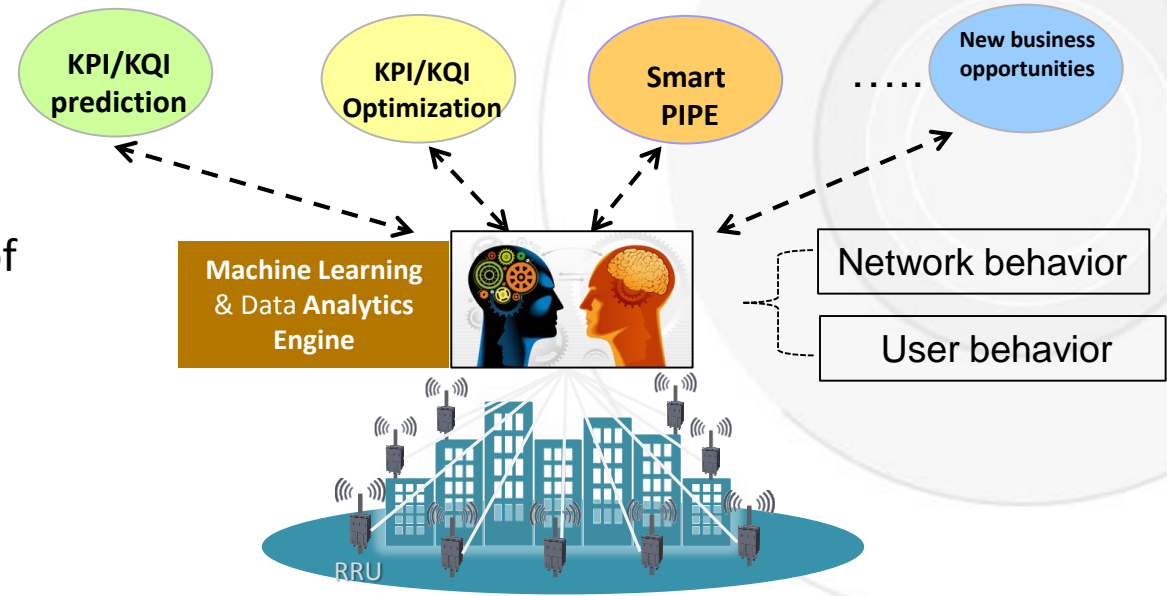


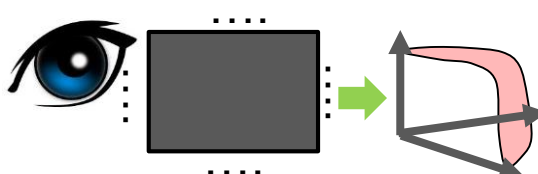


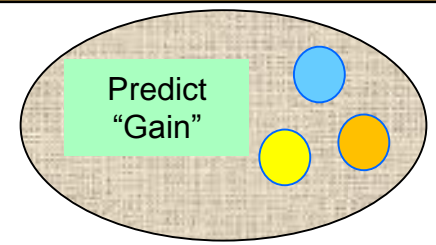
Fig. 1. Sample trajectories of (a) BM, (b) Levy walk, and (c) RWP.



# The future is gene therapy

•How do we spot a new genetic disease in advance? Big data of course!!  
(which is the answer to everything these days...)

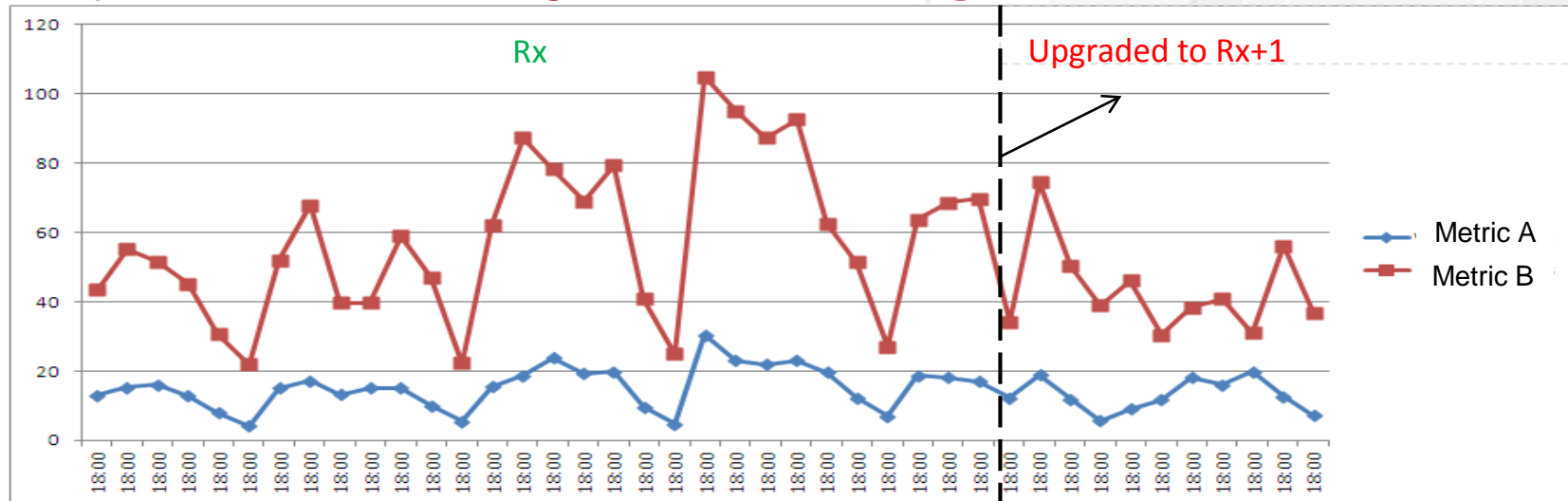


	System modeling	Performance Analysis	Performance Prediction	Benefits
Data Analytics	 <p><b>Black box approach</b></p> <ul style="list-style-type: none"> <li>○ Scale</li> <li>○ Live – self learning/update</li> <li>○ Technology independent</li> <li>○ Need historical data</li> </ul>	 <p><b>What</b></p> <ul style="list-style-type: none"> <li>○ Network-wide, automatic detection</li> <li>○ Identify patterns (where, when, correlations)</li> </ul>	 <p><b>Network wide impacts:</b></p> <ul style="list-style-type: none"> <li>○ Learn the Resource-KPI model</li> <li>○ Predict the KPI changes as a result of new algorithm etc.</li> </ul>	 <p><b>Predictable results:</b></p> <ul style="list-style-type: none"> <li>○ Learn and predict the network wide behavior from the data</li> <li>○ Enhanced operation</li> </ul>

# Example application: Anomaly detection during Software Upgrade

Looks good?

Maybe not...

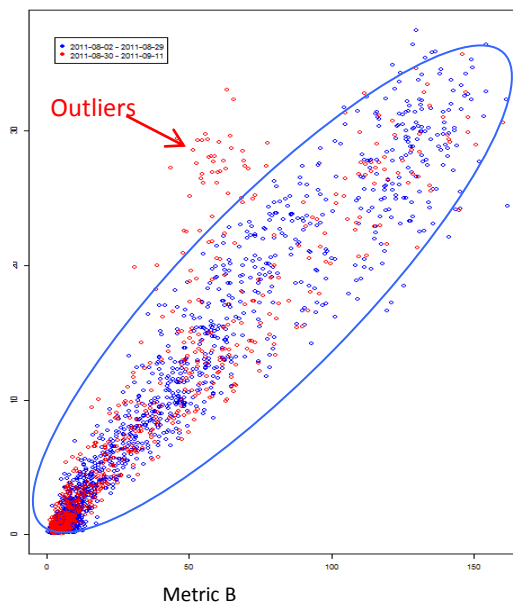
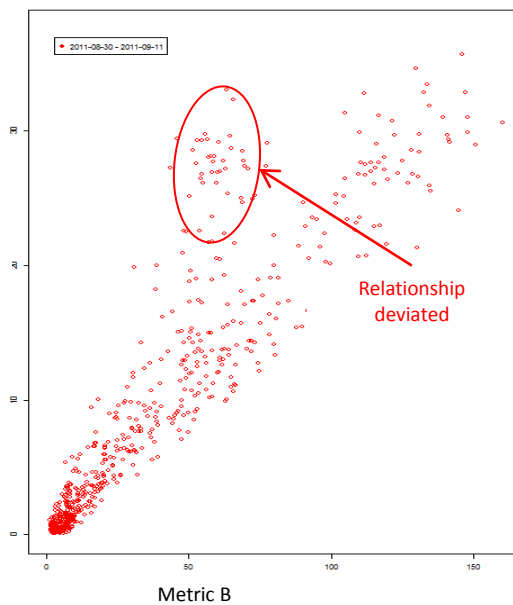
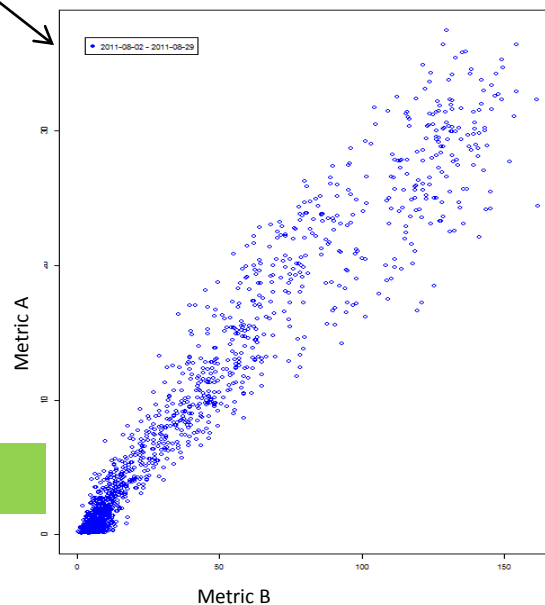


Three key questions and Steps:

Did we introduce **new** bugs?

Did we fix the **known** bugs

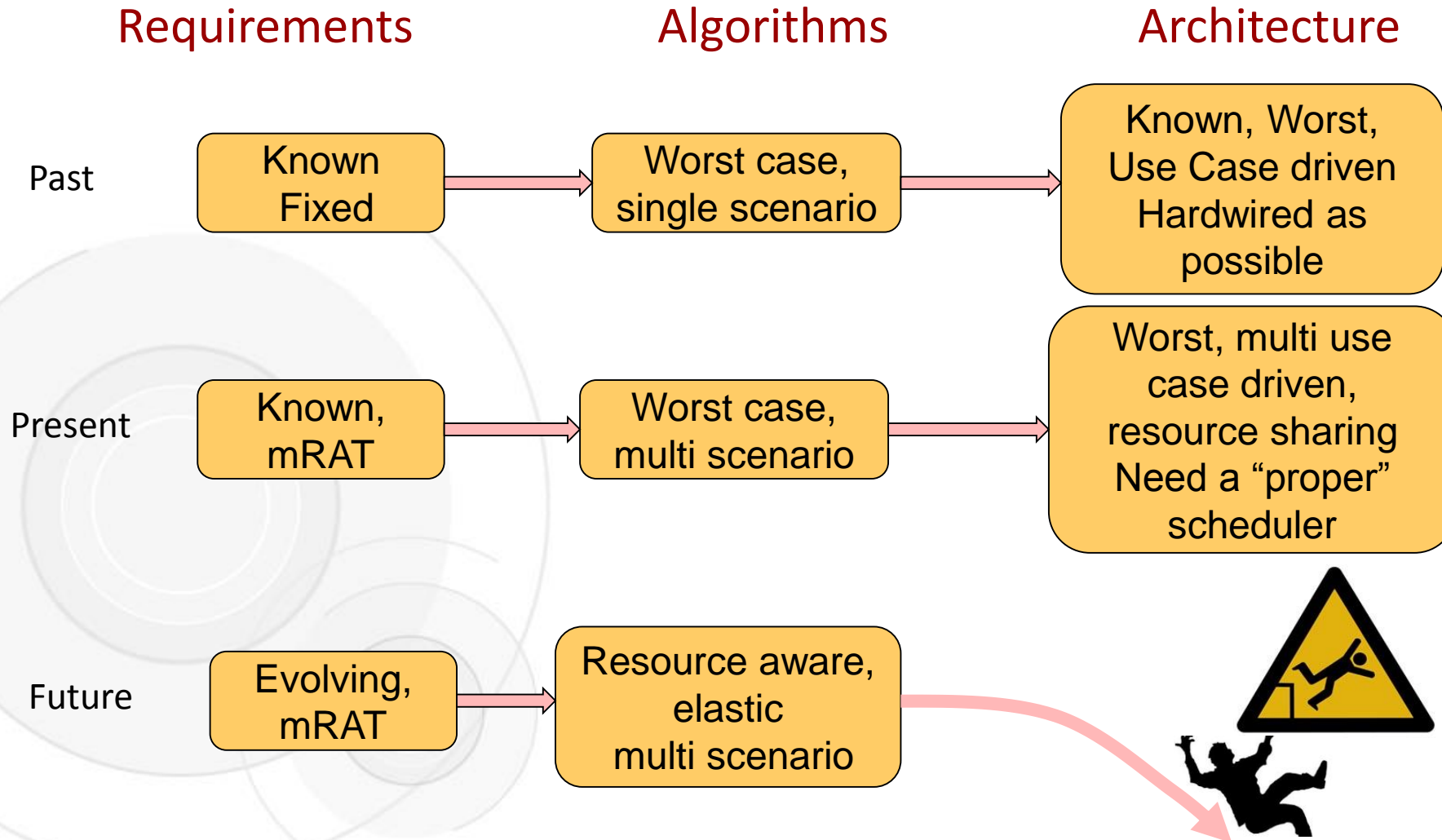
Did we achieve expected **improvements**?



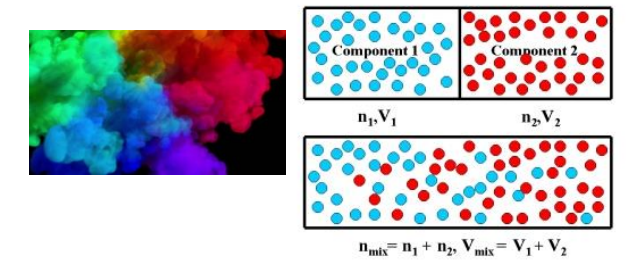
# Part III: Services Run over the Cliff of Real Time

John Milton says I told you so...

# Embedded System Philosophy Hits a Cliff

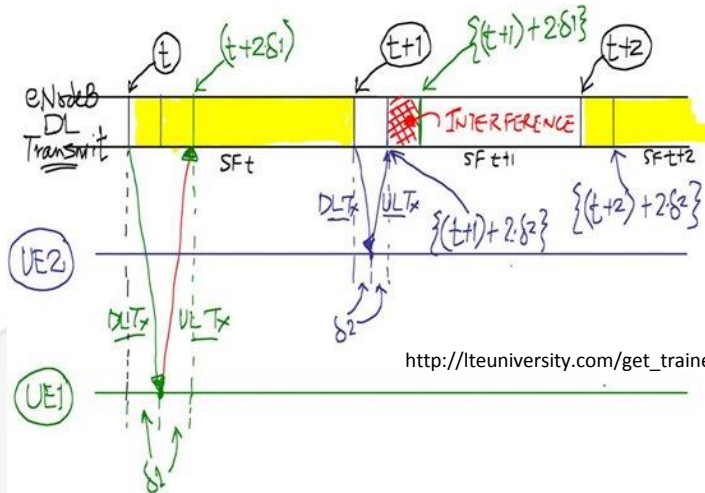


The entropy of mixing for an ideal gas



# From Fixed to Opportunistic: Basic Challenges

## Starting from a Classic Wireless Embedded System



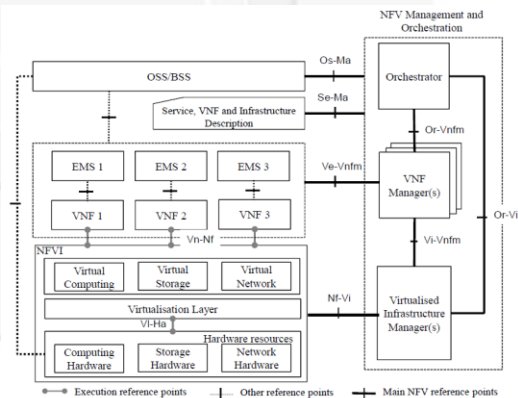
[http://lteuniversity.com/get\\_trained/expert\\_opinion1/b/dhar/archive/2010/08/13/lte-and-the-need-for-time-alignment.aspx](http://lteuniversity.com/get_trained/expert_opinion1/b/dhar/archive/2010/08/13/lte-and-the-need-for-time-alignment.aspx)

Something like this is very confusing from an architecture perspective

Why? There is little apparent **Opportunity to Schedule (OTS)**

But... The high levels of dependency and real time constraints are apparent.

## Starting from a Classic Compute Server Philosophy, NFV etc.



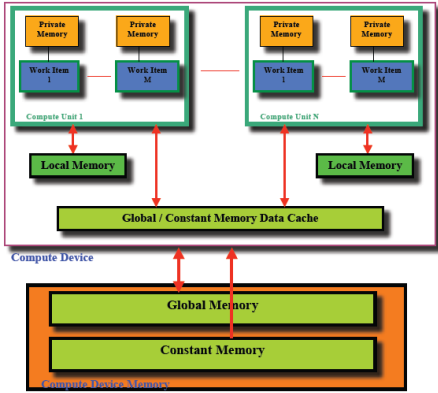
Ignores heterogeneity. Scheduling is spatial (NOVA) and coarse grained

Software overhead is ridiculous

Pretty much ignores scheduling data flow dependencies

# From Fixed to Opportunistic: Basic Challenges

Starting from a GPU approach: DSL is great!! But what DSL?



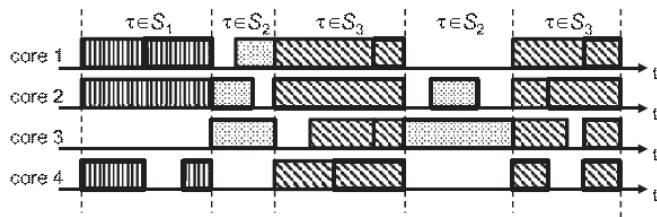
Compute is easy. Data management is hard

Data dependency is easy. Real time constraints is hard.

Probably more than one DSL. Analysis tooling will be hard

**“Innocence, Once Lost, Can Never Be Regained. Darkness, Once Gazed Upon, Can Never Be Lost.”** John Milton. Same is true of real time....

## Starting from a Classic Scheduling Approach



$$\sum_{S_k \in S} L_{j,k} \leq L_j$$

$$\Leftrightarrow \sum_{S_k \in S} \max \left\{ L_j \max_{\tau_{i,k} \in S_k} \{\delta_{i,k}\}, \frac{L_j}{m} \sum_{\tau_{i,k} \in S_k} \delta_{i,k} \right\} \leq L_j$$

$$\Leftrightarrow \sum_{S_k \in S} \max \left\{ \max_{\tau_{i,k} \in S_k} \{\delta_{i,k}\}, \frac{1}{m} \sum_{\tau_{i,k} \in S_k} \delta_{i,k} \right\} \leq 1 \quad (10)$$

*“I have a scheduling policy. Will it support the following real time problem?”*

Scheduling is mostly temporal. Heterogeneity now being addressed

Schedulability continues to be mostly worst case. Leads to unacceptably loose bounds

Tend to focus on closed form solutions, but this is changing

Network Scheduling is statistical, but needs a lot of adaptation

# Some Promising Approaches

Measurement based probabilistic timing analysis

## Probabilistic Timing Analysis on Time-Randomized Platforms for the Space Domain

Mikel Fernandez<sup>†</sup>, David Morales<sup>‡</sup>, Leonidas Kosmidis<sup>‡</sup>, Alen Bardizbanyan<sup>\*</sup>, Ian Broster<sup>‡</sup>, Carles Hernandez<sup>‡</sup>, Eduardo Quinones<sup>‡</sup>, Jaime Abella<sup>‡</sup>, Francisco Cazorla<sup>†,‡</sup>, Paulo Machado<sup>‡</sup>, Luca Fossati<sup>¶</sup>  
<sup>†</sup>Barcelona Supercomputing Center (BSC) <sup>‡</sup>Rapita Systems LTD  
<sup>§</sup>Spanish National Research Council (IIIA-CSIC) <sup>¶</sup>European Space Agency <sup>\*</sup>Cobham Gaisler

MBTA to MBPTA. Data mine the crap out of it and build a model  
Need enough data mined from the SoC. This is a challenge

Bayesian Theory applied to scheduling.

Need enough processors and jobs to create a statistic  
Don't schedule. Randomize and play the statistics

## Scheduling Storms and Streams in the Cloud

Javad Ghaderi  
Columbia University  
New York, NY  
jghaderi@columbia.edu

Sanjay Shakkottai  
University of Texas  
Austin, TX  
shakkottai@austin.utexas.edu

R Srikant  
University of Illinois  
Urbana, IL  
rsrikant@illinois.edu

## Random Modulo: a New Processor Cache Design for Real-Time Critical Systems

Carles Hernandez<sup>‡</sup>, Jaime Abella<sup>‡</sup>, Andrea Gianarro<sup>‡</sup>, Jan Andersson<sup>‡</sup>, Francisco J. Cazorla<sup>†,‡</sup>  
<sup>†</sup>Barcelona Supercomputing Center (BSC-CNS), Barcelona (Spain)  
<sup>‡</sup>Cobham Gaisler, Gothenburg (Sweden)  
<sup>\*</sup>Spanish National Research Council (IIIA-CSIC), Barcelona (Spain)

## Network Slicing Games: Enabling Customization in Multi-Tenant Mobile Networks

Pablo Caballero<sup>\*</sup> Albert Banchs<sup>†</sup> Gustavo de Veciana<sup>\*</sup> Xavier Costa-Pérez<sup>‡</sup>  
<sup>\*</sup>The University of Texas at Austin, Austin, TX. Email: pablo.caballero@utexas.edu, Gustavo@ece.utexas.edu  
<sup>†</sup>University Carlos III of Madrid and IMDEA Networks Institute, Madrid, Spain. Email: banchs@it.uc3m.es  
<sup>‡</sup>NEC Laboratories Europe, Heidelberg, Germany. Email: xavier.costa@neclab.eu

More generally connecting L2 and L1 scheduling

Consider the network statistics when scheduling L1 resources

## Fog-RAN: Hardware Resource Sharing in Aggregated Baseband Processing Systems

Huishan Zhu  
The University of Texas at Austin  
huishanz@utexas.edu

Alan Gatherer  
Huawei Technologies Co Ltd  
alan.gatherer@huawei.com

Mattan Szter  
The University of Texas at Austin  
mattan.szter@mail.utexas.edu

- OTS: You need many different ways to do the same thing
- Scheduling Granularity: Not too coarse, not too fine...
- OTS must produce more benefit than it loses due to “softening” of processing

# Meanwhile...Compute Server Design Also Evolves: Machine Learn Everything



Machine Learning for Systems  
and  
Systems for Machine Learning

Jeff Dean  
Google Brain team  
[g.co/brain](http://g.co/brain)

## Device Placement with Reinforcement Learning

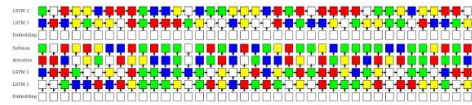
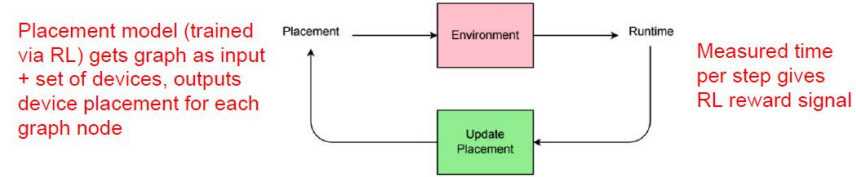


Figure 4. RL-based placement of Neural MT graph. Above: encoder. Below: decoder. Devices are denoted by colors, where the transparent color represents an operation on a CPU and each other unique color represents a different GPU. This placement achieves an improvement of 19.3% in running time compared to the fine-tuned hand-crafted placement.

+19.3% faster vs. expert human for neural translation model

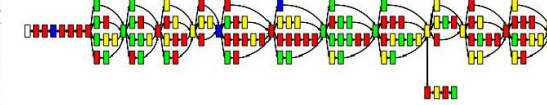


Figure 5. RL-based placement of Inception-V3. Devices are denoted by colors, where the transparent color represents an operation on a CPU and each other unique color represents a different GPU. RL-based placement achieves the improvement of 19.7% in running time compared to expert-designed placement.

+19.7% faster vs. expert human for InceptionV3 image model

If you can automate device placement in the cloud you can automate SoC mapping and even scheduling

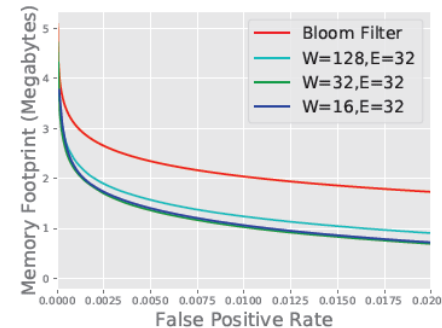


Figure 13: Learned Bloom filter improves memory footprint at a wide range of false positive rates. (Here  $W$  is the RNN width and  $E$  is the embedding size for each character.)

If you can learn a better hash filter and reduce area by 50% you can do that for caches too



# Conclusions

- Compute Servers vs baseband: getting harder to tell the difference
  - They can learn from each other. Trick is to achieve real time and service orientation
- 5G is the most exciting standard ever
  - Service focused. Many new players and new ideas
- The service is the customer, the human is the food, we are the restaurant
  - We need to start seeing the patterns hidden in the network
  - These patterns change from location to location and service to service and over time
  - Yes of course ML will play a role....
- Real Time Support of Services is an unsolved problem
  - It will become more and more boutique in space and time.
  - How to adapt scheduling schemes? How to test??
  - Yes of course ML will play a role....

# Shameless Plug for the CTN

- Alan Gatherer, Editor-in-Chief, [comsoc.org/ctn](http://comsoc.org/ctn)
- Hottest Topics Delivered Monthly: 72K Opt-in Subscribers (as of 2016)

- Provides a quirkier, volunteer version of IEEE Spectrum, focused specifically on Communications Issues
- Tries to engage experts in hot topics and get a “read over coffee” level of article with references for the reader who is interested in further education.
- Publication via push email to the website once a month



## The Challenges of 5G in a Cloud Based Network

IEEE CTN Issue: April 2018

This month we consider some of the challenges that face the deployment of 5G. In particular we look at deployment on cloud based networks and the challenges of network slicing 5G in this environment. Anwer, Shahid and Qiang provide us with a comprehensive picture of the pieces that will have to fall into place before we can get to the promised land of a truly software defined and sliced 5G network. As always, your comments are most welcome.

[Full Version](#)

[Leave a comment](#)

## Machine Learning Modems: How ML will change how we specify and design next generation communication systems

IEEE CTN Issue: March 2018

So, I hear you all ask, are we all out of a job? Are communication engineers another victim of the rise of the machine? Well the answer in this article is thankfully no, not yet. But Nathan, Ben and Tim point us towards a fascinating new way to specify and design communication systems that may change forever the way we standardize, design and field optimize our products. Time to take that “basics of ML” online course you have been putting off for the last year! But first, read this article. Comments, and recommendations for good classes in ML, welcome as always in the comments section at the end of the article. Alan Gatherer EIC, CTN.

[Full Version](#)

ADVERTISEMENT

### FLUOROWRAP® Electrical Insulation for Wire and Cable

Saint-Gobain's FLUOROWRAP® solutions include fluoropolymer tapes and heat sealable polyimide-fluoropolymer composites that are engineered for the most demanding electrical insulation.

ADVERTISEMENT

### IEEE members save up to 35% with UPS, available for new or existing shippers.

ADVERTISEMENT

### Attend Now: On Demand Webinar

#### Choosing the Right Probing Solution and Oscilloscope for Power Electronics Measurements

Sponsored by Rohde & Schwarz

## Join Our Community!

- Neutral Hosting: A Piece of the 5G Puzzle?** by Alan Gatherer
- DEATH BY STARVATION? BACKHAUL AND 5G**
- 5G, SOCIAL JUSTICE AND THE ROLE OF THE IEEE COMSOC**
- 5G AND THE NEXT BILLION MOBILE USERS: A VIEW FROM AFRICA**
- 5G AND THE NEXT BILLION MOBILE USERS: A VIEW FROM INDIA**
- Lost in Space: How Secure Is The Future For Mobile Positioning?**
- 5G And The Next Billion Mobile Users: A View From Africa**
- Is Anyone Out There? 5G, Rural Coverage And The Next 1 Billion**
- THE DEATH OF 5G (PART 2)**
- Resurrection of 5G In Defense of Massive MIMO**

# Thanks!

[www.huawei.com](http://www.huawei.com)

HUAWEI TECHNOLOGIES CO., LTD.

