what if our computers could see?
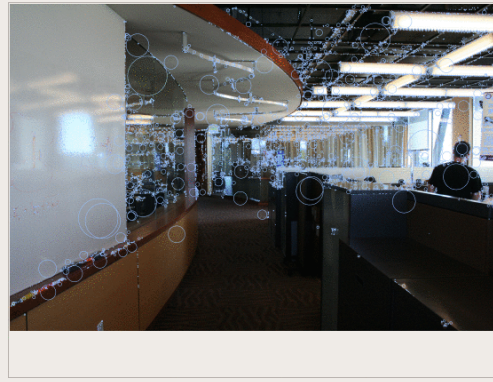
Microsoft's's HoloLens

who?

john

where?

what?

red apple
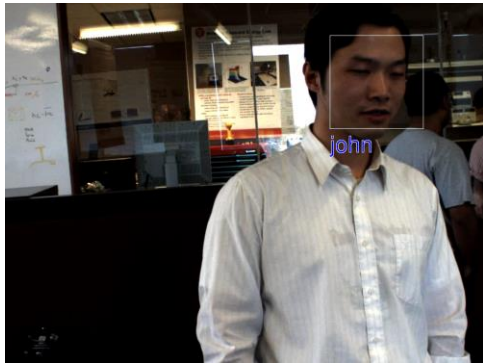
Microsoft
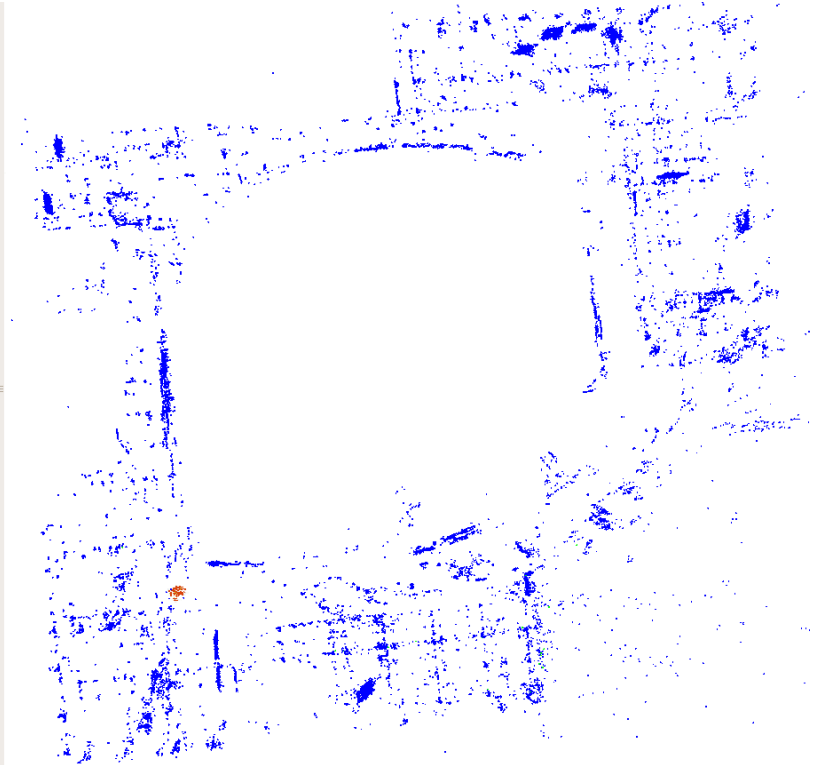
seeing is for real

Since February of 2012, Rialto, California has required all police officers to wear a camera to monitor all interactions with the public.

After this practice was instituted, the number of complaints to the department dropped by 88%. Further, the number of instances of the police using "force" on people dropped 60%.

"When you put a camera on a police officer, they tend to behave a little better, follow the rules a little better..."

– Chief William Farrar

FB/POLICETHEPOLICEACP

BBC NEWS LONDON updated at 11:11 ET

Metropolitan Police officers start wearing body cameras

The New York Times

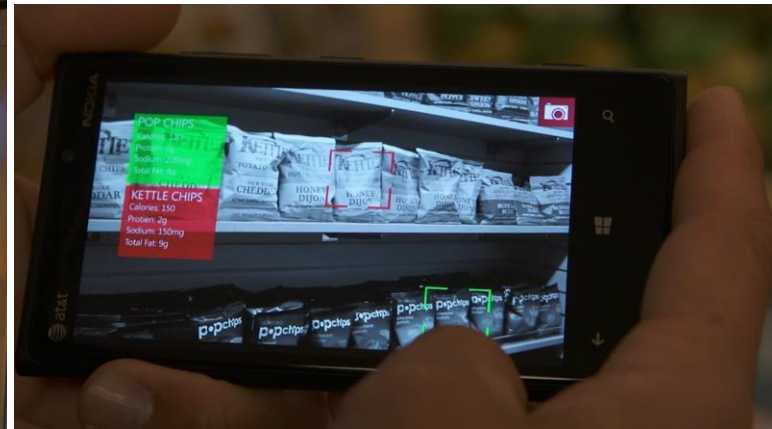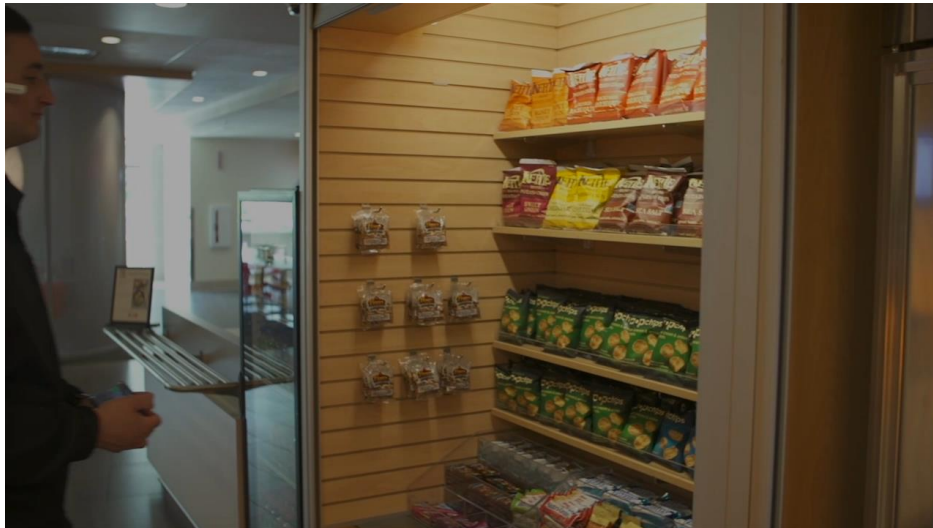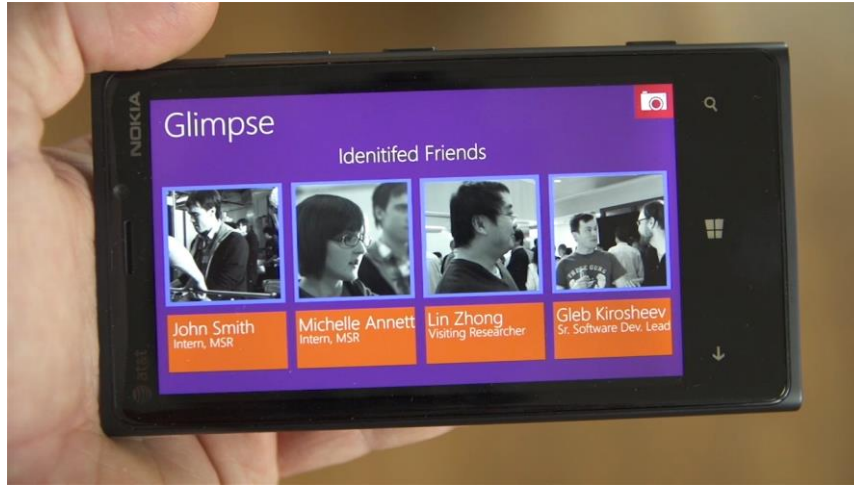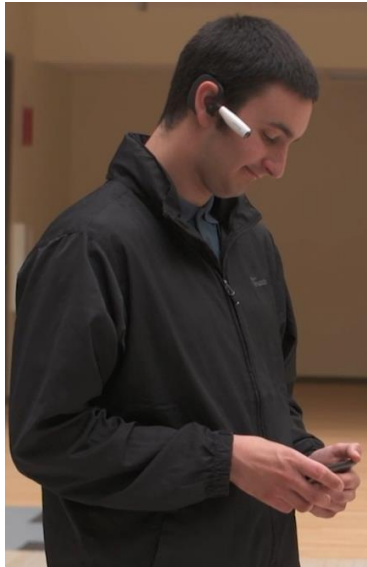*New York Police Officers to Start Using Body Cameras in a Pilot Program*

By J. DAVID GOODMAN   SEPT. 4, 2014

The Washington Post

D.C. police will wear body cameras as part of pilot program

Microsoft

MSR's Glimpse project

vision is demanding

recognition using *deep neural networks*

	face[1] [1]	scene [2]	object[2][3]
memory (floats)	103M	76M	138M
compute	1.00 GFLOPs	2.54 GFLOPs	30.9 GFLOPs
accuracy	97%	51%	94% (top 5)

1: 4000 people;   2: 1000 objects from *ImageNet*, *top 5:* one of your top 5 matches

> human-level accuracy, heavy resource demands
> … offloading computation is highly desirable

[1] Y. Taigman *et al*. ***DeepFace: Closing the Gap to Human-Level Performance in Face Verification***. In CVPR 2014. **(Facebook)**

[2] B. Zhou *et al*. ***Learning deep features for scene recognition using places database***. In NIPS, 2014. **[MIT, Princeton, ..]**

[3] K. Simonyan & A. Zisserman. ***Very deep convolutional networks for large-scale image recognition***. 2014 **[Google, Oxford]**

Microsoft

recognition: server versus mobile

Object Detection | Feature Extraction | Object Recognition/Labeling
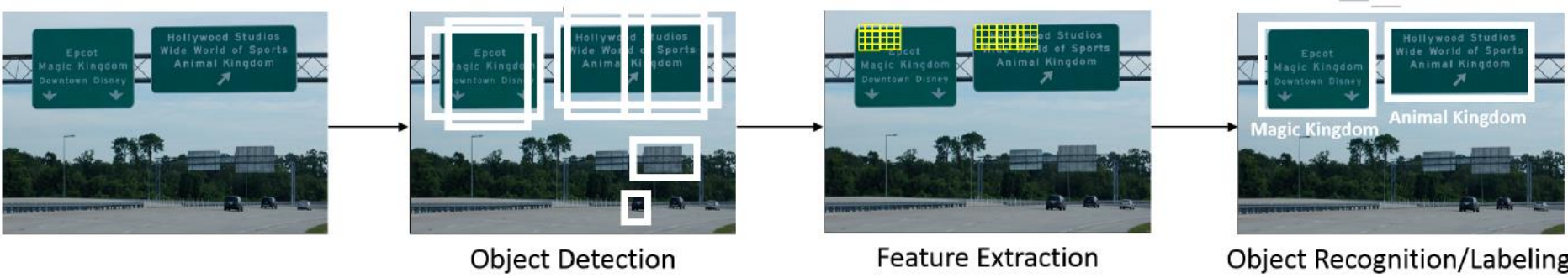
road sign recognition[1]

stage	Mobile (Samsung Galaxy Nexus)	server (i7, 3.6GHz, 4-core)	Spedup (server:mobile)
detection	2353 +/- 242.4 ms	110 +/- 32.1 ms	**~15-16X**
feature extraction	1327.7 +/- 102.4 ms	69 +/- 15.2 ms	**~18X**
recognition[2]	162.1 +/- 73.2 ms	11 +/- 1.6 ms	**~14X**
Energy used	11.32 Joules	0.54 Joules	**~21X**

[1]convolution neural networks
[2]classifying 1000 objects with 4096 features using a linear SVM

Microsoft

how long does it take to reach the cloud?

3g networks

T-Mobile 450ms

AT&T 350ms
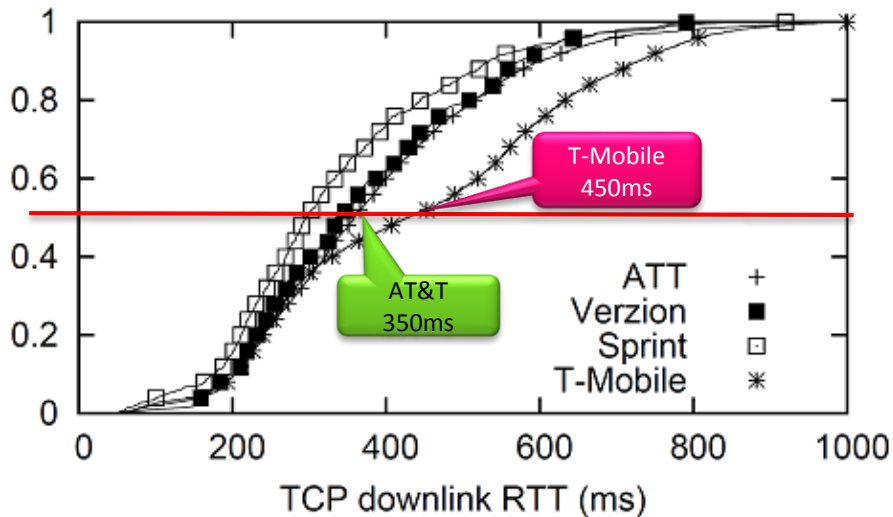
MobiSys 2010

4g-lte networks

MobiSys 2013

2 years later, we still have latency issues
(May 9, 2015)

major cloud provider A

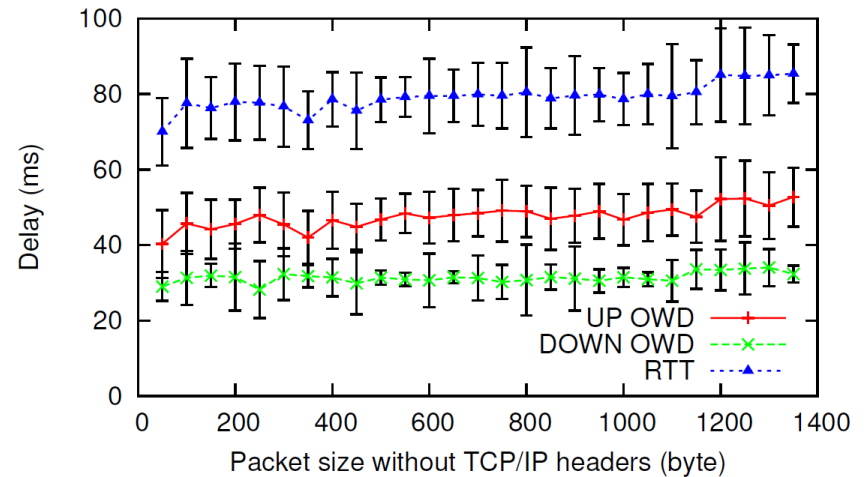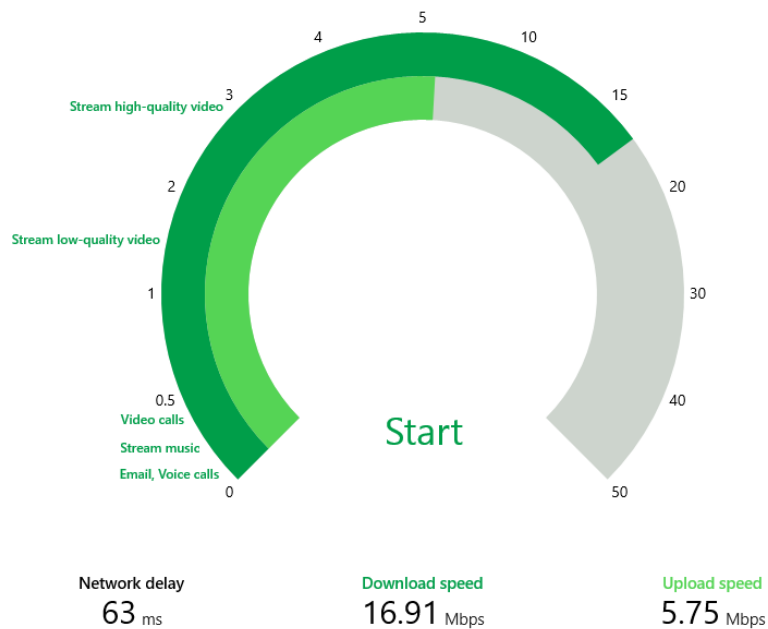| Data Center	Average Latency
West US	115ms
South Central US	131ms
East US	155ms
North Central US	171ms
North Europe	222ms
West Europe	223ms
Japan West	251ms
East Asia	251ms
Japan East	253ms
Southeast Asia	253ms
Central US	276ms
Content Delivery Network *	276ms
East US 2	287ms
Brazil South	371ms
Australia Southeast	398ms
Australia East	441ms

major cloud provider B

```
Command Prompt

C:\Users\bahl>tracert -d 209.85.225.99

Tracing route to 209.85.225.99 over a maximum of 30 hops

  1    38 ms    25 ms    47 ms  172.26.96.169
  2    45 ms    39 ms    29 ms  172.18.84.36
  3   109 ms    39 ms    39 ms  12.249.2.25
  4    59 ms    88 ms    70 ms  12.83.180.2
  5    81 ms    71 ms    88 ms  12.122.31.194
  6    76 ms    72 ms    87 ms  12.122.136.181
  7     *         *         *     Request timed out.
  8   106 ms    62 ms    80 ms  216.239.49.168
  9    81 ms   100 ms   111 ms  209.85.246.253
 10    90 ms   124 ms   112 ms  216.239.46.212
 11   110 ms   119 ms   119 ms  72.14.239.48
 12   135 ms   135 ms   133 ms  209.85.243.99
 13   138 ms   120 ms   121 ms  216.239.46.214
 14   120 ms   120 ms   125 ms  209.85.249.213
 15   192 ms   119 ms   123 ms  209.85.247.6
 16   116 ms   126 ms   112 ms  64.233.175.45
 17     *       ^C
C:\Users\bahl>
```

also, http://claudit.feld.cvut.cz/claudit/rtdata.php

Microsoft

try it out –
download Microsoft's Network Speed Test

Network Speed Test

Last Test (2/12/2013 12:50 PM)

Stream high-quality video
Stream low-quality video
Video calls
Stream music
Email, Voice calls

Start

Current network

Connection type
Wi-Fi
Network name
A-MSFTWLAN
Internet status
Internet access
Host name
minint-6d
Access point BSSID
6C:F3:7F:4F:88:72
Authentication
Rsna
Encryption
Ccmp

Settings

Network Speed Test
By Microsoft Research

Options

About

Privacy Statement

Contact us

Permissions
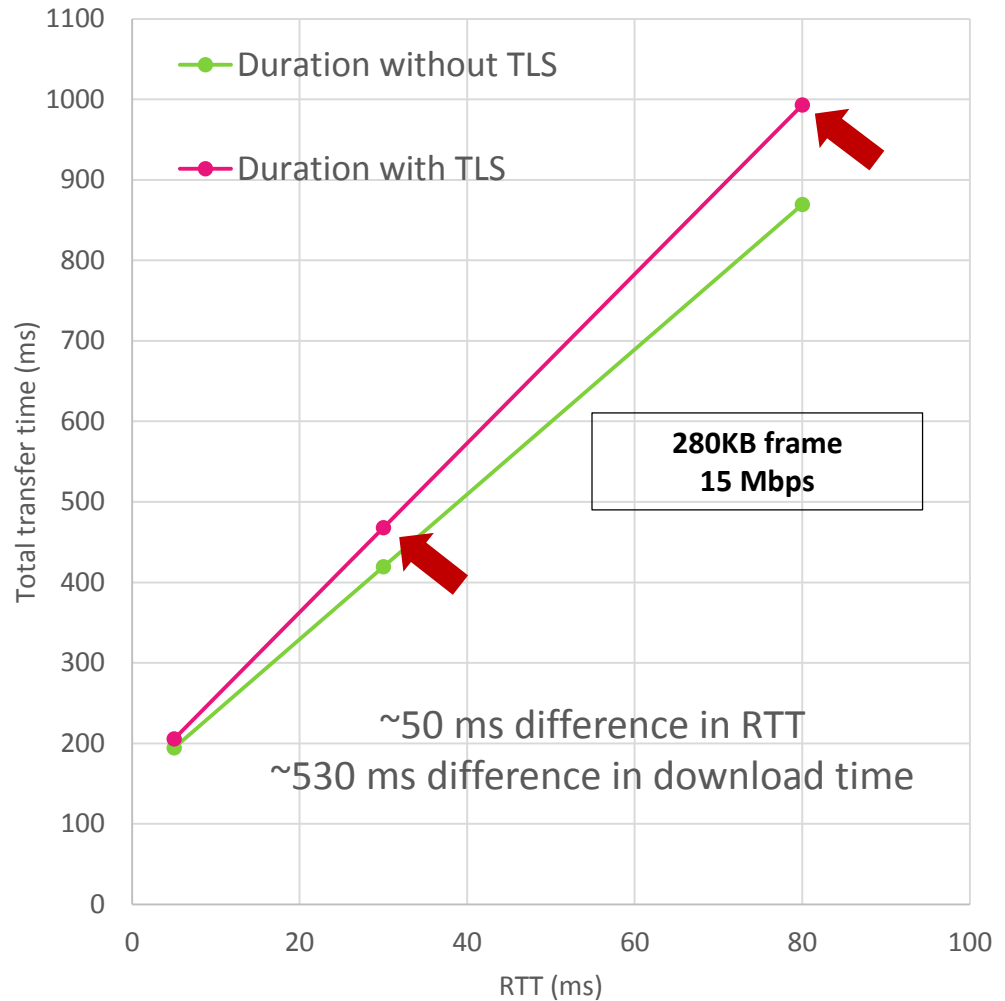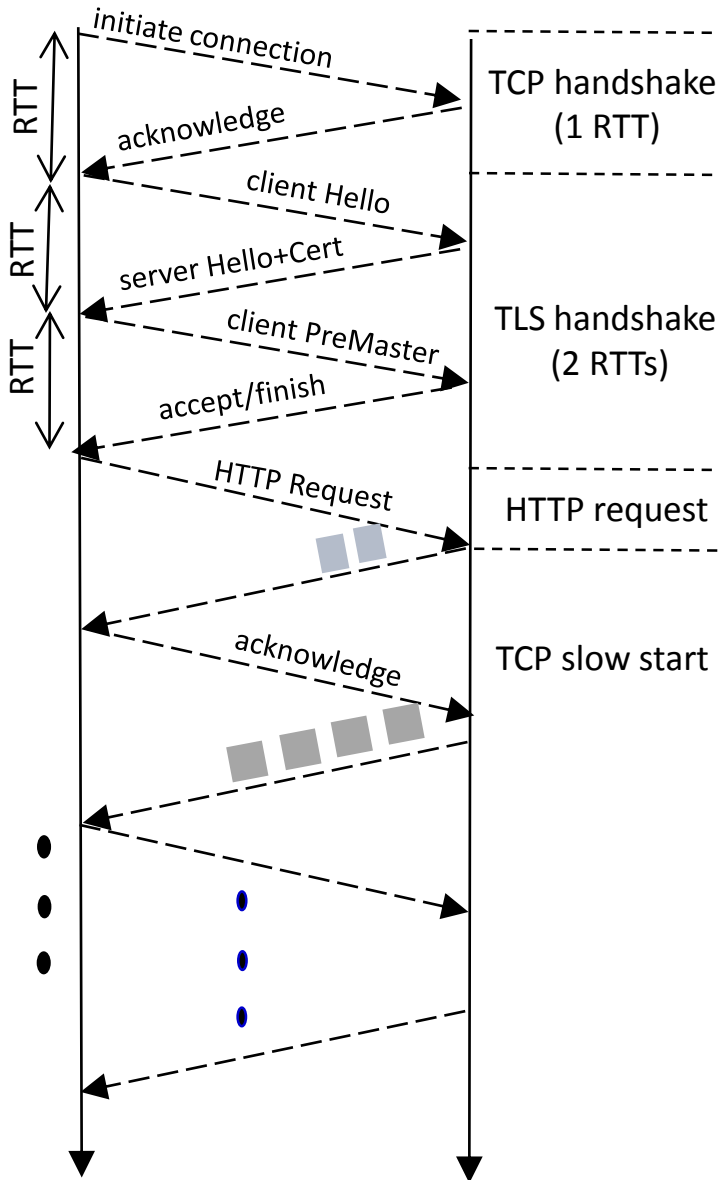
A-MSFTWLAN   98   Screen

Notifications   Power   Keyboard

Change PC settings

Network delay
63 ms

Download speed
16.91 Mbps

Upload speed
5.75 Mbps

Available on Windows Phone and Windows 8

Microsoft

popular protocols make things worse!

RTT — initiate connection → TCP handshake (1 RTT)
RTT — acknowledge ←
RTT — client Hello → TLS handshake (2 RTTs)
RTT — server Hello+Cert ←
client PreMaster →
accept/finish ←
HTTP Request → HTTP request
acknowledge ← TCP slow start

Duration without TLS
Duration with TLS

280KB frame
15 Mbps

~50 ms difference in RTT
~530 ms difference in download time

Total transfer time (ms) vs RTT (ms)

even with UDP - end user impact
fast action cloud gaming

client

server

Client Input

→ network →

Process Input

Update Game State

Render Frame

H264 Encode

← network ←

H264 Decode

Show Frame

example:
- RTT to server = 80ms (LTE)

input lag:

16 capture input
+40 one-way net delay
+16 primary processing
+20 encode delay
+40 one-way net delay
+10 decode & display

Total: ~140 ms

Microsoft

11
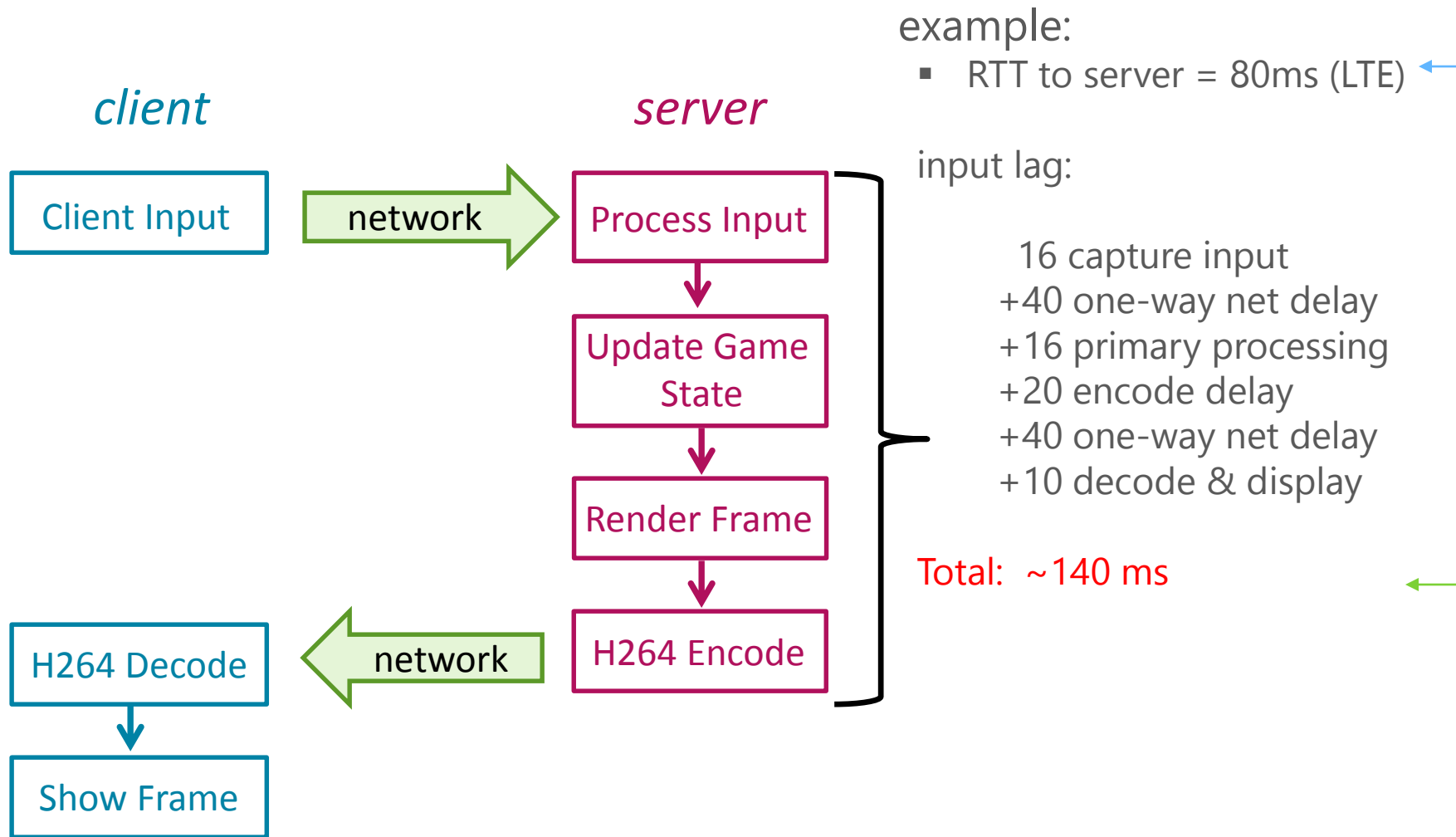
impact of 5, 30 & 80 msec latency
(fast action gaming)

Microsoft
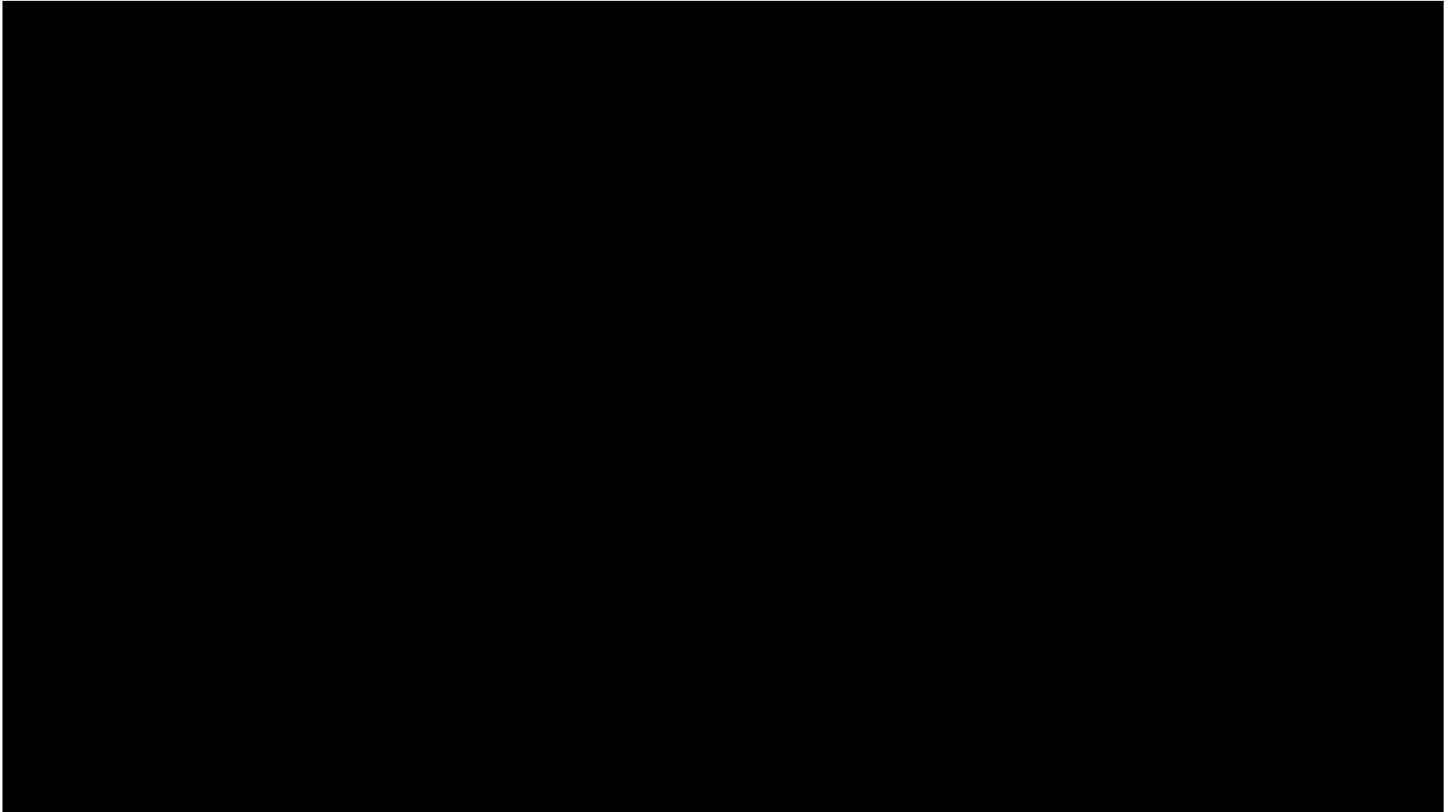
latency matters!

"being fast really matters...half a second delay caused a 20% drop in traffic. and it killed user satisfaction"

- Marissa Mayer @ Web 2.0 (2008)

"...a 400 millisecond delay resulted in a -0.59% change in searches/user",
[i.e. Google *would lose 8 million searches per day - they'd serve up many millions fewer online adverts*]

- Jake Brutlag, Google Search (2009)

"...for Amazon every 100 ms increase in load times decreased sales with 1%"

- Andy King, book author

"...when 50% of traffic was redirected to our edges preliminary results showed a 5.9% increase in click-thru rates"

- Andy Lientz, Partner GPM, BingEdge (2013)

Microsoft

the fact of the matter is ...

offloading computation to a resource-rich cloud brings the true power of CS into your hands
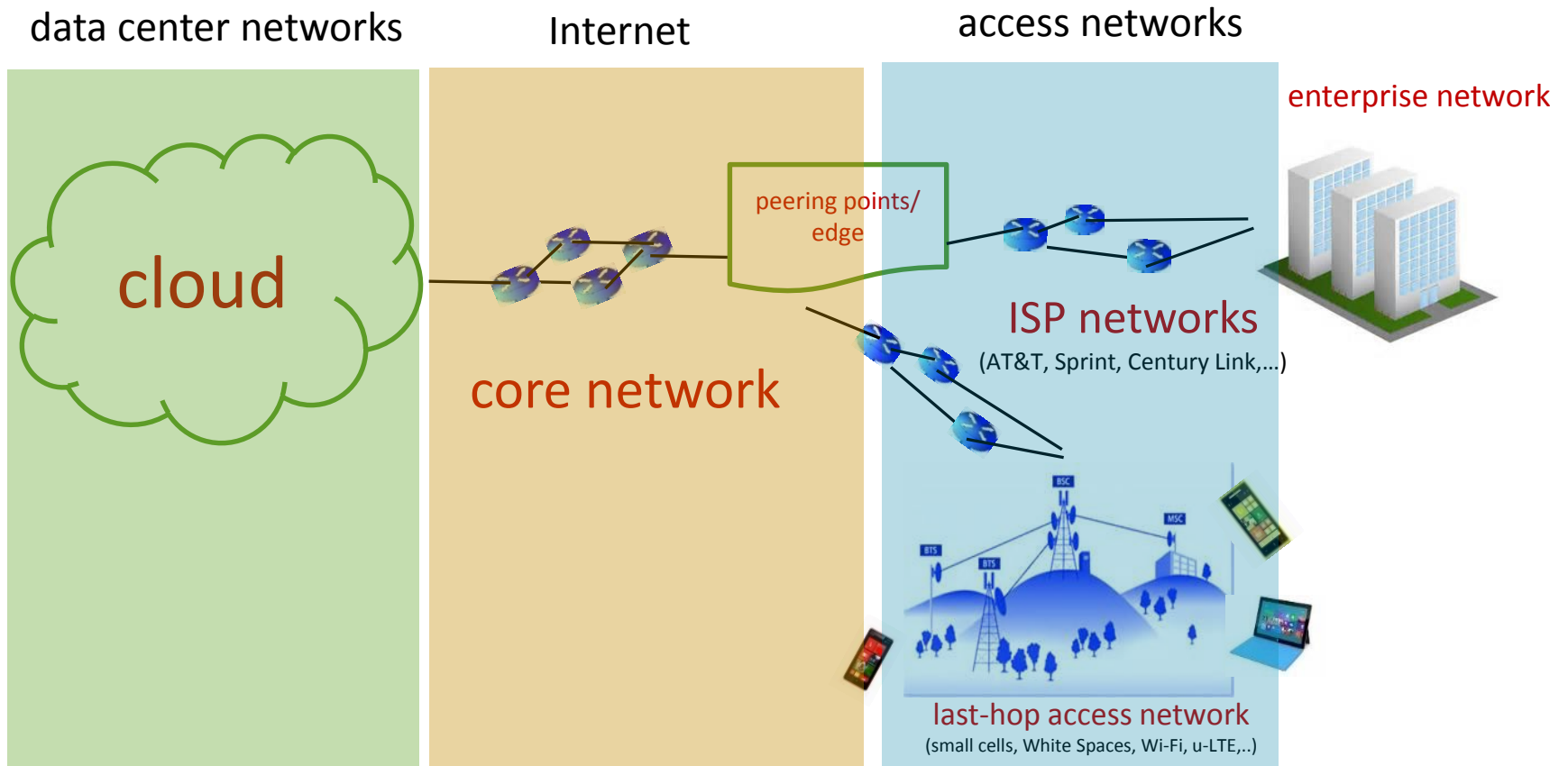
high latency & jitter to the cloud can make cloud services unusable

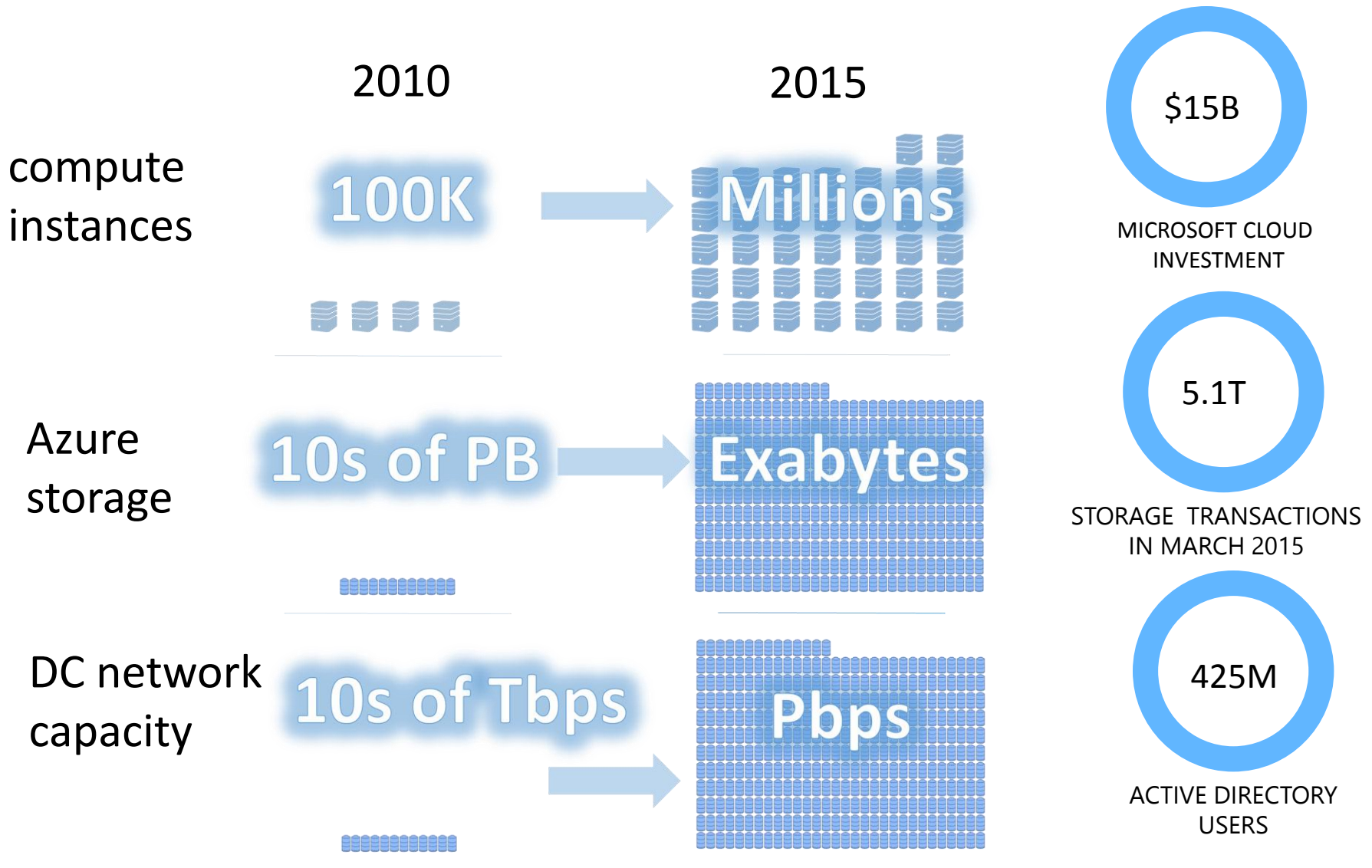poor performance impacts revenue and turns users away

... and we have a latency problem

Microsoft

reducing latency

contributors to latency
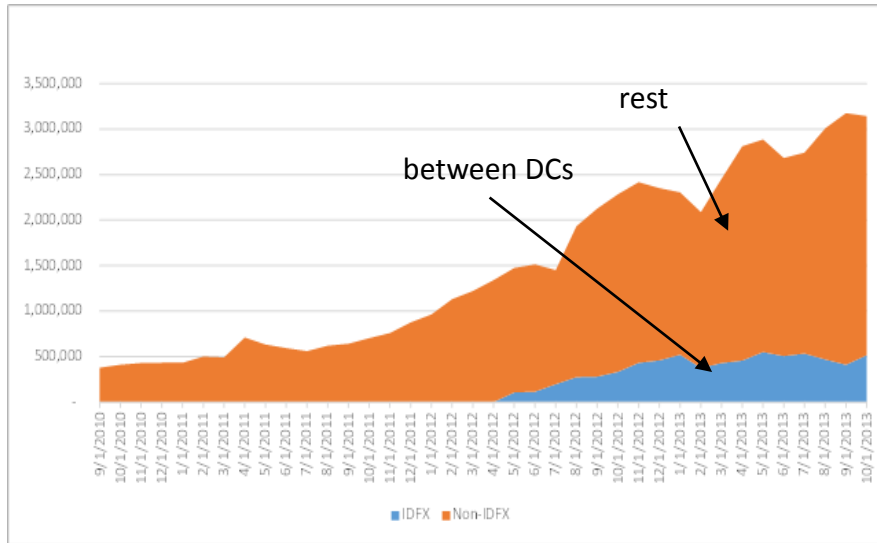

data center networks

Internet

access networks

enterprise network

cloud

peering points/edge

core network

ISP networks
(AT&T, Sprint, Century Link,…)

last-hop access network
(small cells, White Spaces, Wi-Fi, u-LTE,..)

Microsoft

Microsoft's hyper-scale cloud

2010   2015

compute instances

100K → **Millions**

$15B

MICROSOFT CLOUD INVESTMENT

Azure storage

10s of PB → **Exabytes**

5.1T

STORAGE TRANSACTIONS IN MARCH 2015

DC network capacity

10s of Tbps → **Pbps**

425M

ACTIVE DIRECTORY USERS

Microsoft

Microsoft's hyper-scale network

Microsoft's network is one of the largest in the world

1.4M

MILES OF FIBER
(DC & WAN)

4X

WRAP THE EARTH IN
NORTH AMERICAN FIBER

massive traffic growth is stressing the underlying core networks

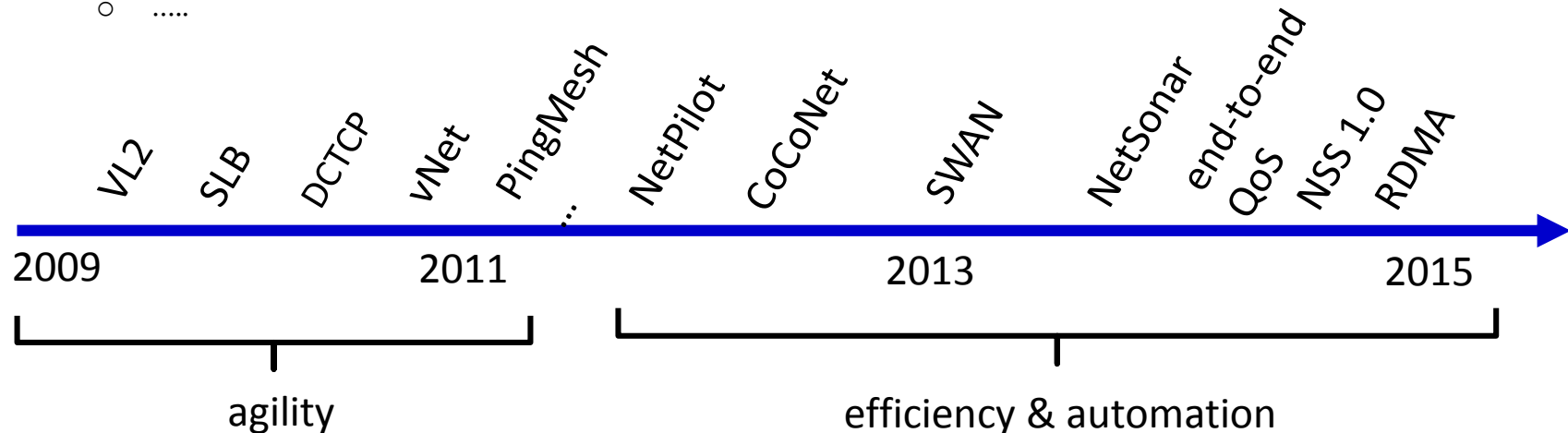areas MSR researchers are working on: **SIGCOM 2014**

performance: significant number of circuits sit idle while others are oversubscribed (latency increases)

failures: long convergence time during network topology changes with planned and unplanned network events

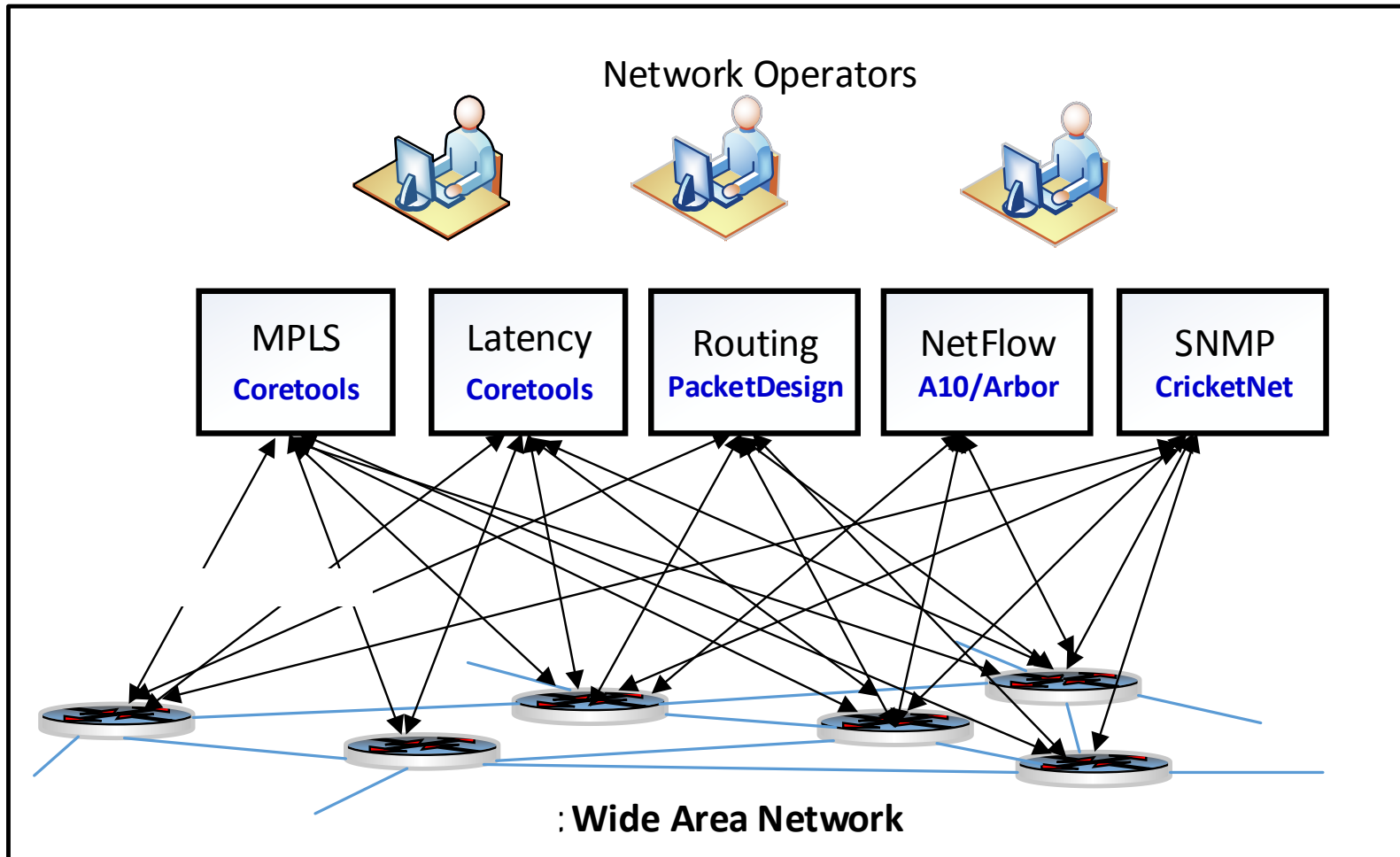# MSR's contributions to Microsoft cloud networking & to academia

reseachers worked hand-in-hand with Azure, Bing, Windows, ...

- steady stream of significant tech transfers
  - o full –bisection bandwidth (Q10): 80x cost reduction, 20x outage reduction, in all Azure DCs
  - o software load balancer (SLB): 15x cost reduction, carries all Azure traffic
  - o software-defined WAN: increased inter-DC bw utilization from ~40% to ~95%,
  - o virtual networking: enabled MSFT hybrid cloud offering via HyperV virtual network product
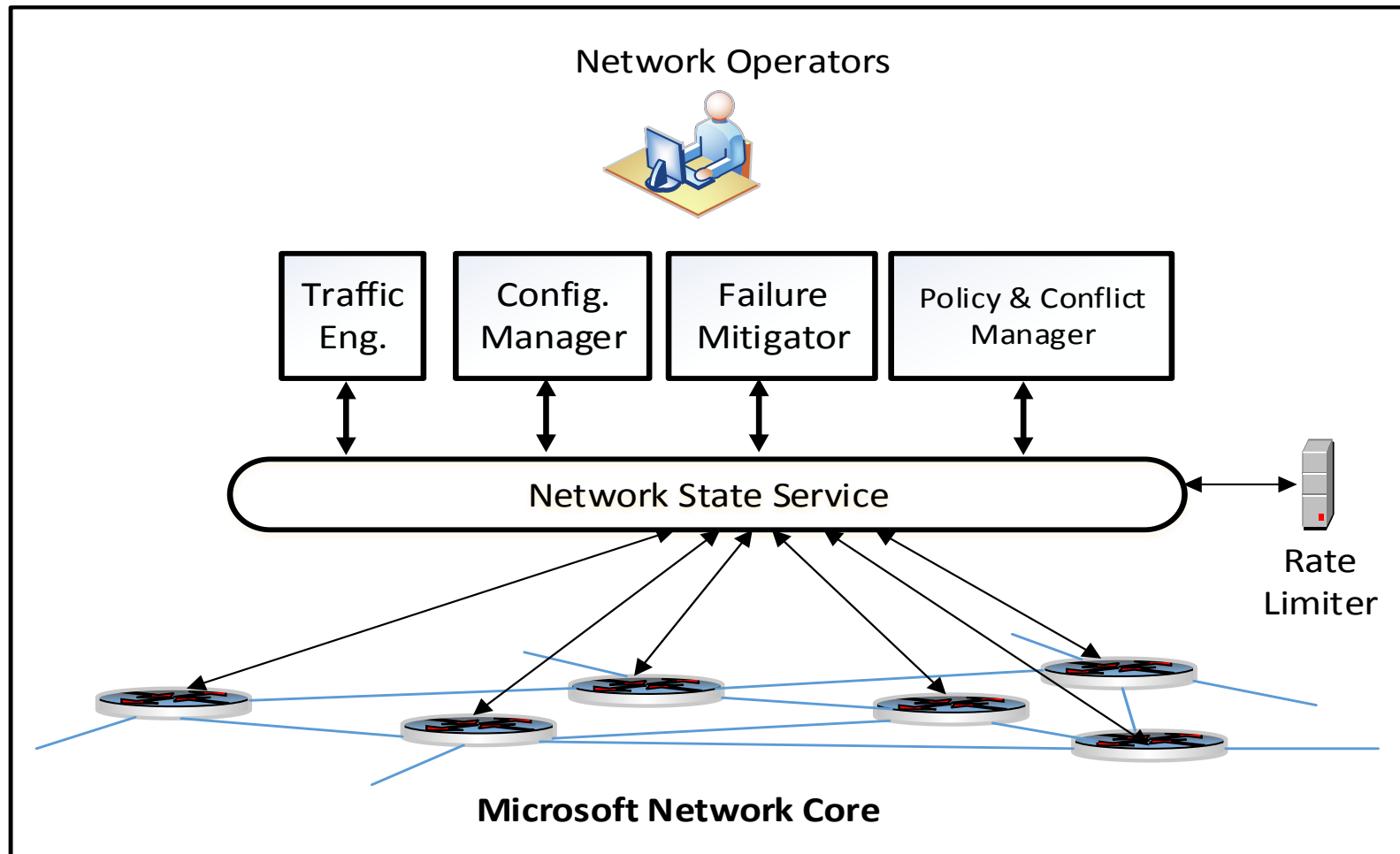 - o .....

Timeline markers: VL2, SLB, DCTCP, vNet, PingMesh, NetPilot, CoCoNet, SWAN, NetSonar, end-to-end QoS, NSS 1.0, RDMA

2009 — 2011 — 2013 — 2015

agility | efficiency & automation

- plenty of research accolades as well
  - o papers recognized as "Research Highlight" by ACM
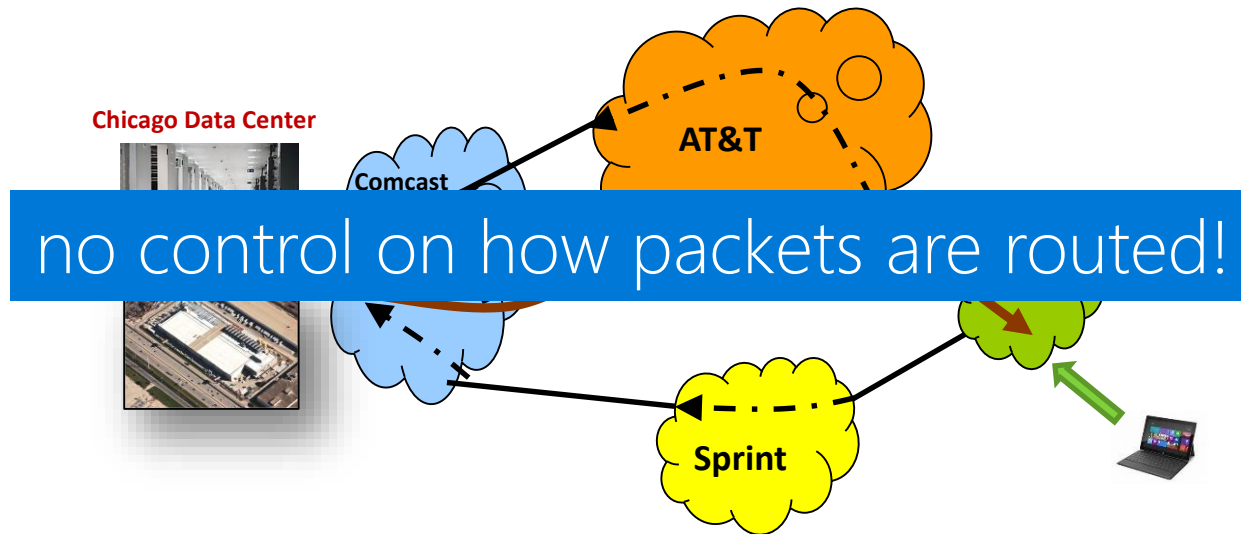
Microsoft

improving efficiency of wide area network

Network Operators

| MPLS **Coretools** | Latency **Coretools** | Routing **PacketDesign** | NetFlow **A10/Arbor** | SNMP **CricketNet** |

: **Wide Area Network**

improving efficiency of wide area network
with MSR's network state service

Network Operators

| Traffic Eng. | Config. Manager | Failure Mitigator | Policy & Conflict Manager |

Network State Service

Rate Limiter

Microsoft Network Core

Internet: a network of networks of networks

a collection of many autonomous systems (AS) managed by many ISPs with complex peering relationships

Chicago Data Center

Comcast

AT&T

no control on how packets are routed!

Sprint

as of March 6, 2013 (source: PEER 1)
- 22,961 AS numbers (AS numbers uniquely identify networks on the Internet, e.g. 8075 for Microsoft)
- 50,519 peering connections

Microsoft
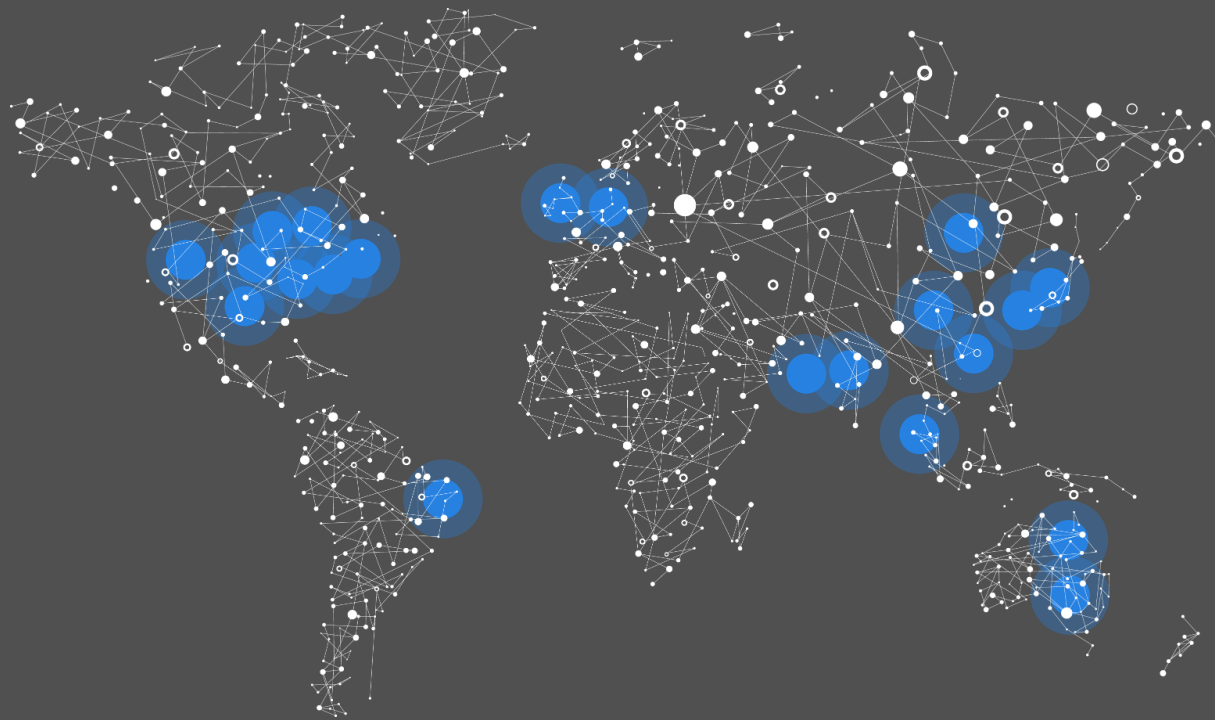
... but we can reduce latency further

get the packets under our control as soon as possible

how?

- bring the <u>cloud closer</u> to the end-user
  - ✓ build lots of DCs around the world & place them in strategic locations

Microsoft

bringing the cloud closer
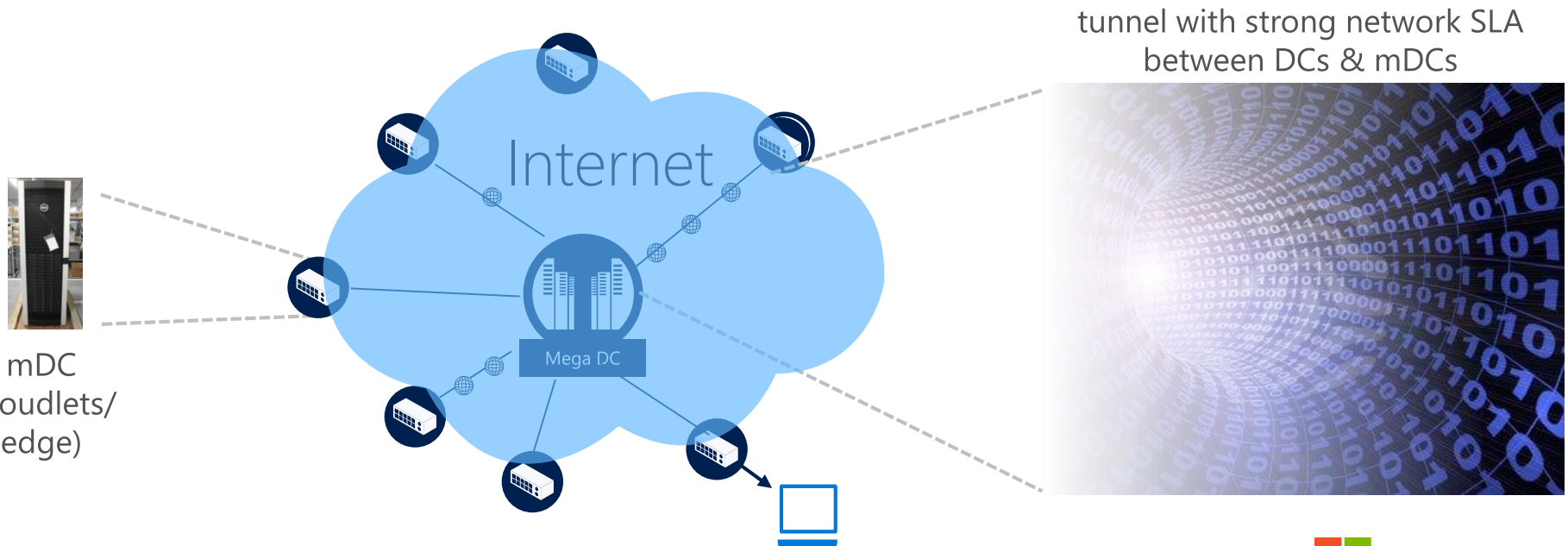## build lots of hyper-scale data centers around the world

19

Azure compute regions open today

more than AWS and
Google cloud combined

● Azure compute regions

is building hyper-scale data centers enough?

no, it's capital intensive and expensive to operate

smarter approach: build an extensive infrastructure of micro DCs (1-10s of servers with several TBs of storage, $20K-$200K/mDC) and place them everywhere

tunnel with strong network SLA between DCs & mDCs

Internet

Mega DC

mDC
(cloudlets/
edge)

Microsoft

micro DCs
site acceleration (classic)

content caching
- o Xbox videos, NetFlix videos, Windows updates,...

split TCP connections
- o from Bing data, on avg. can reduce latencies by ~30 msec
  - o predictive search query responses improved ~25-35% based on random sampling before and after deploying edge serves in a couple of US cities

Akamai
Limelight
CloudFront
Level 3
EdgeCast
Rackspace
⋮
⋮

mDCs are "classic" CDNs nodes, that can improve the performance of search engines, office productivity tools, video and audio conferencing & future cloud services

Microsoft

additional benefits of mDCs

latency reduction

- ✓ serve static content immediately
- ✓ SSL termination / split TCP
- ▪ edge to DC protocol enhancements

bandwidth saving

- ✓ compression
- ▪ procrastination
- ▪ edge analytics

service & internet monitoring

reliable connectivity

- ▪ overlay networking
- ▪ path diversity

battery saving

- ▪ computation offloads
- ▪ client proxying

high-end game streaming
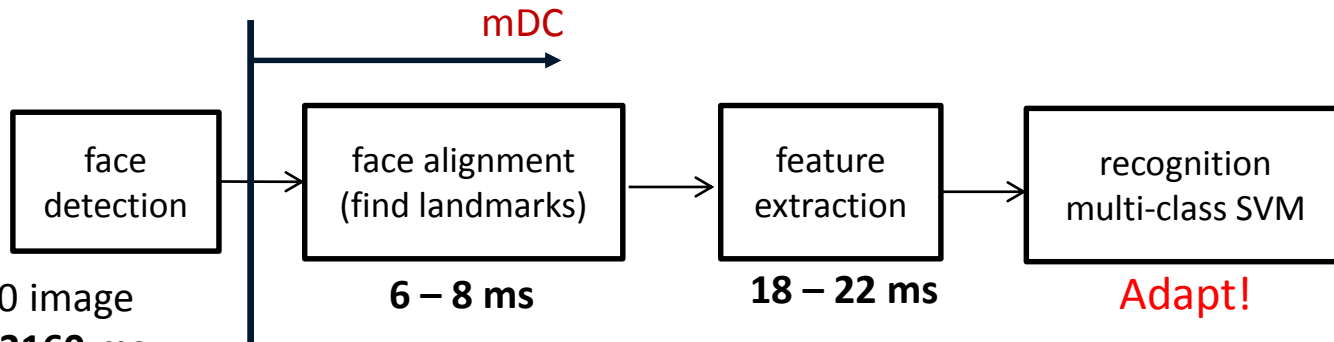
- ▪ lower device cost
- ▪ reduce developer fragmentation
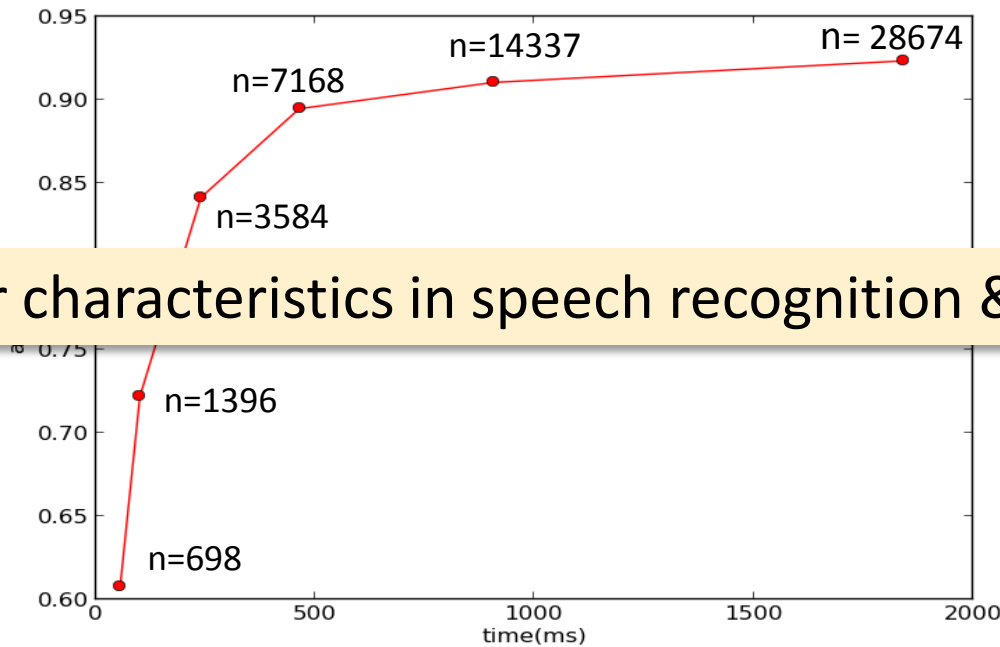
new services

protection against DoS
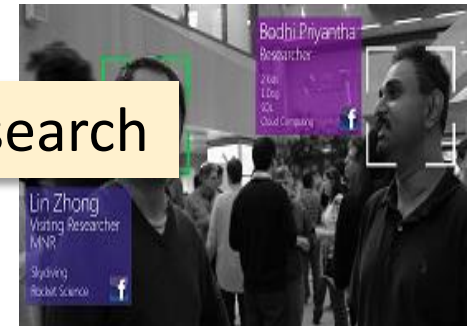
reduced load on DCs

new services: object recognition
the lower the latency, the better the results

mDC

face detection	face alignment (find landmarks)	feature extraction	recognition multi-class SVM
	6 – 8 ms	**18 – 22 ms**	Adapt!

For a 640x480 image
client: 890 – 3160 ms
server: 72 – 115 ms

n= 28674

n=14337

n=7168

n=3584

n=1396

n=698

Face prediction Time

time(ms)

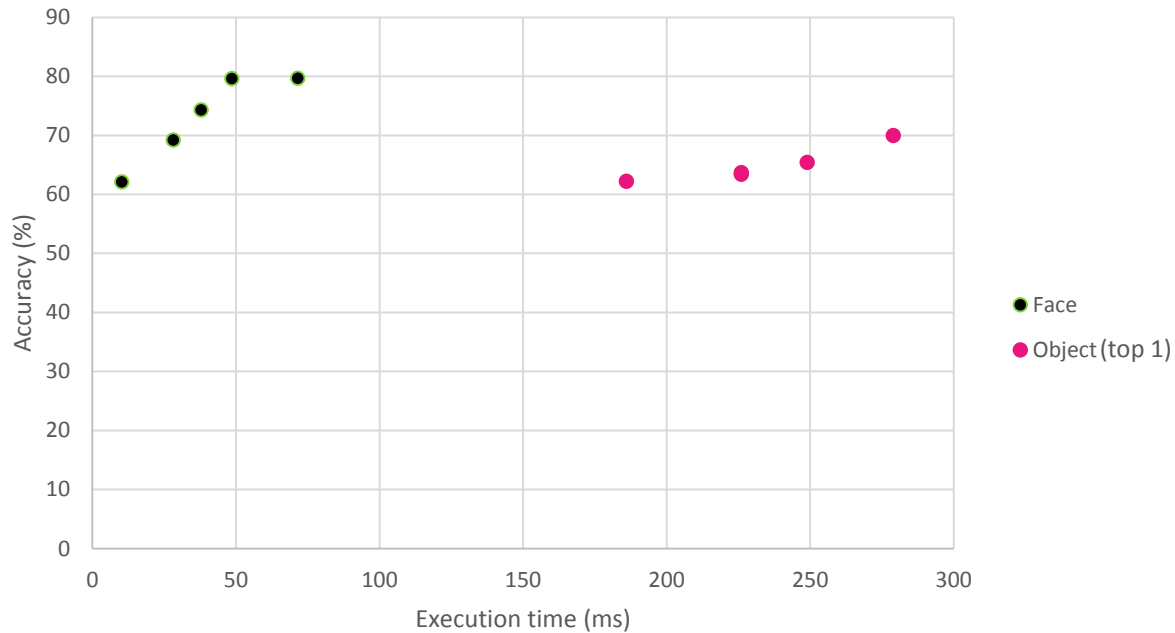similar characteristics in speech recognition & search

Microsoft

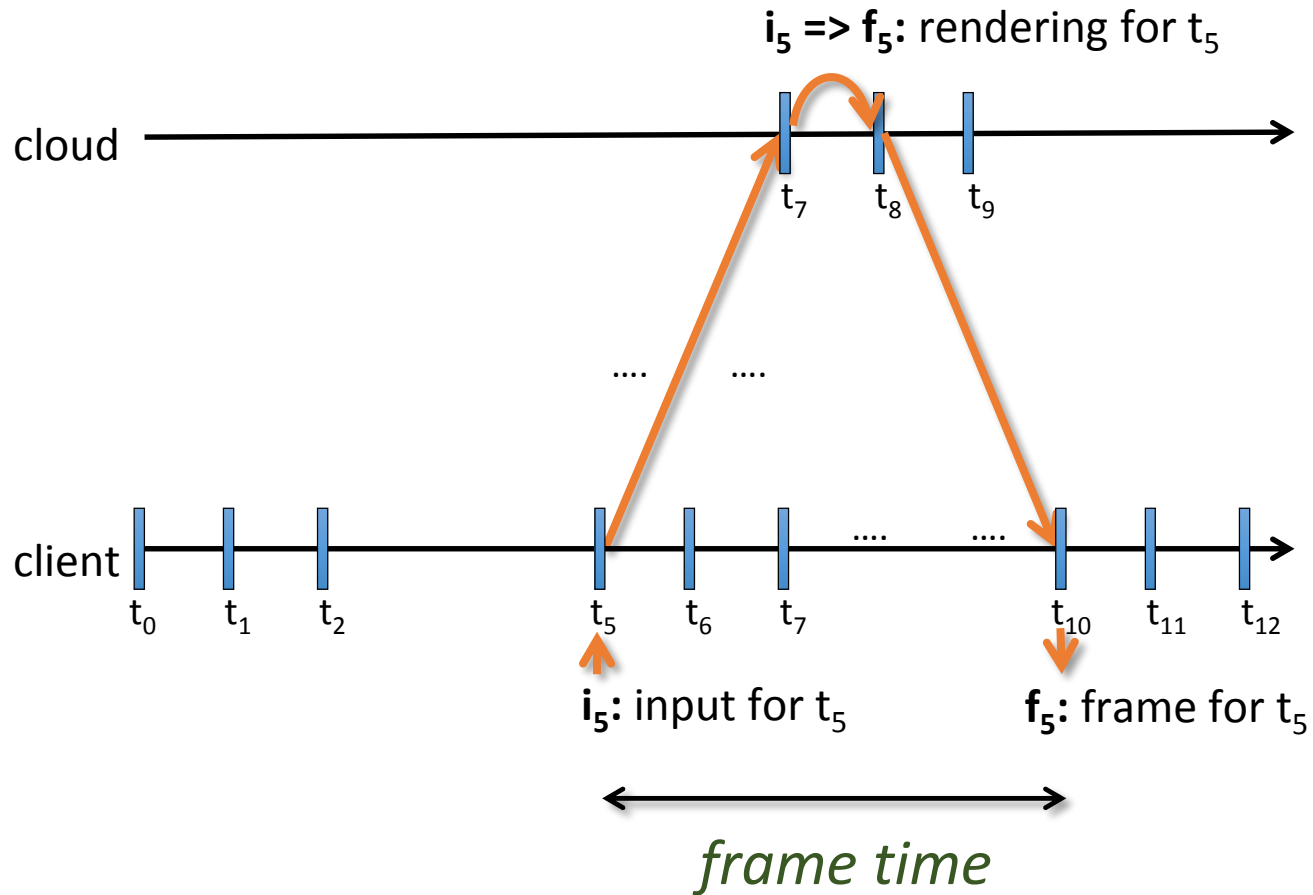# using DNNs - similar results - lower transport latency helps

model execution time vs. accuracy (core i7)

50-100ms can allows ~10-20% more accurate model

face recognition with mDCs

(new) service: cloud gaming

cloud gaming
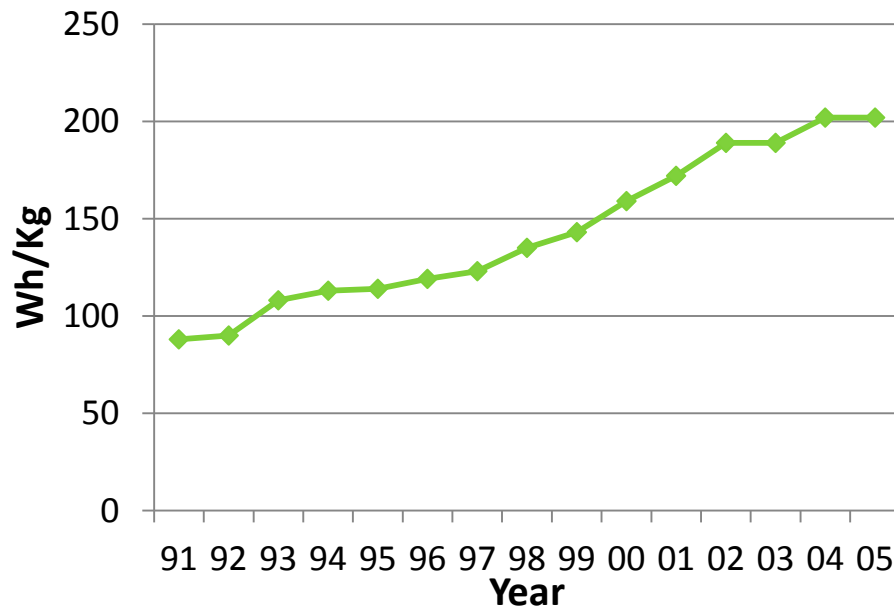## (with speculative execution)

battery life…
silver bullet seems unlikely

Li-Ion energy density

lagged behind
- higher voltage batteries (4.35 V vs. 4.2V) – 8% improvement
- silicon anode adoption (vs. graphite) – 30% improvement

trade-offs
- fast charging = lower capacity
- slow charging = higher capacity

contrast with
CPU performance improvement during same period: 246x

Microsoft

battery use in SmartPhones...

LTE consumes > 1.5W when active
LTE chip active for ~10 secs of extra tail time (1W power)

Verizon numbers from 2012

Power (W)

1.6
1.0

DATA

0.01

T   T + 0.1   T + 11

Time (in seconds)
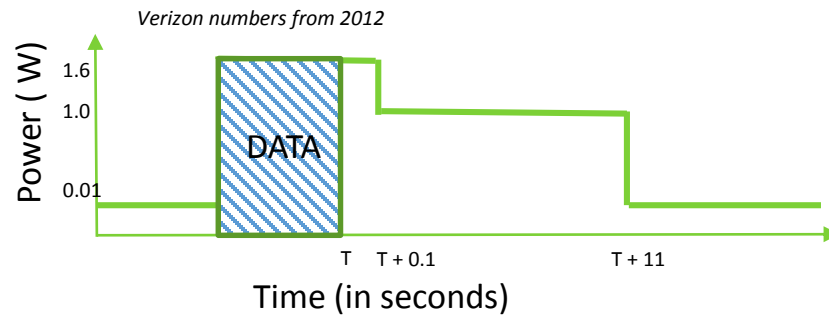
....but how did we get here

Microsoft

a bit of context/history... 6 years ago

The New York Times

Customers Angered as iPhones Overload AT&T

By JENNA WORTHAM
Published: September 2, 2009

The New York Times

DIGITAL DOMAIN
AT&T Takes the Blame, Even for the iPhone's Faults

By RANDALL STROSS
Published: December 12, 2009
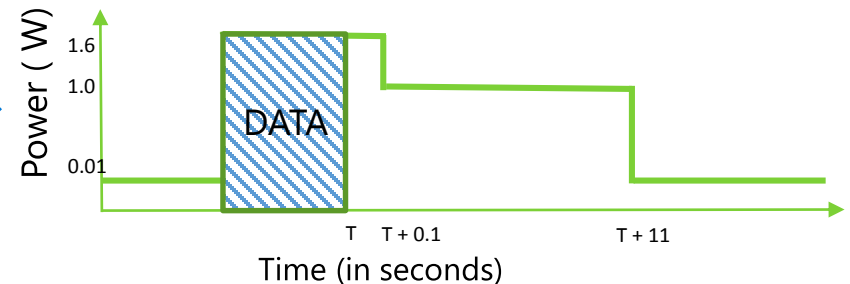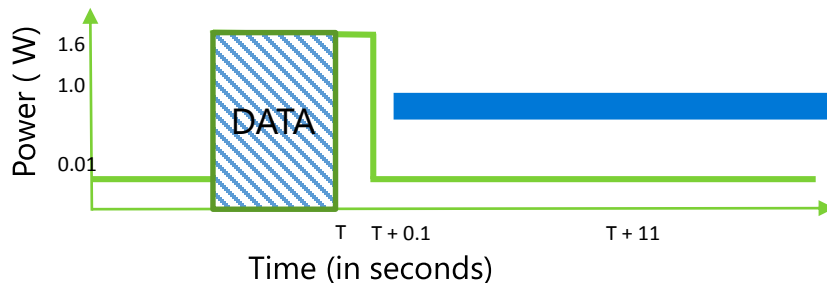
PCWorld

Report: AT&T Reputation Tarnished by iPhone Flaws

By Tony Bradley, PCWorld   Dec 14, 2009 2:01 PM

original design:
bring radio to low power state immediately

mobile operator requirement:
keep LTE chip **active for ~10 sec**. of extra tail time
(to reduce the signaling load)

Verizon numbers from 2012

Left graph: Power (W) vs Time (in seconds). Values: 1.6, 1.0, 0.01. DATA block between T and T + 0.1. Markers at T, T + 0.1, T + 11.

Right graph: Power (W) vs Time (in seconds). Values: 1.6, 1.0, 0.01. DATA block between T and T + 0.1. Markers at T, T + 0.1, T + 11.

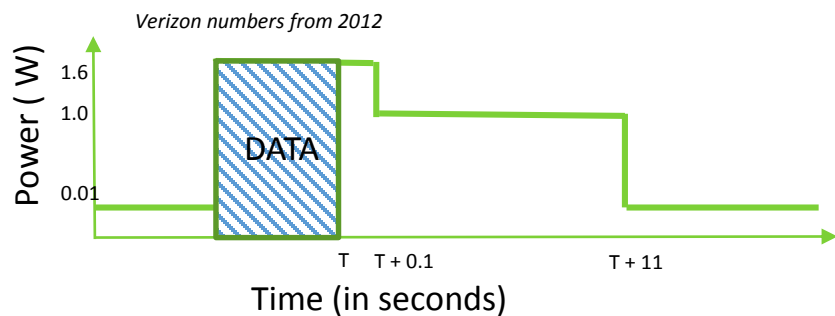**Microsoft**

mDCs can increase use time

LTE consumes > 1.5W when active
LTE chip active for ~10 secs of extra tail time (1W power)

with mDCs:
faster transfers => less time in high power state
aggressively enter lowest power state

Verizon numbers from 2012
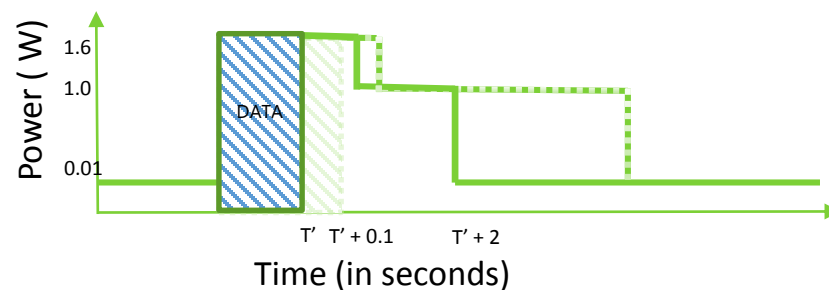
Energy savings / transfer: 1.6W*speedup + 1W*9sec = 10.6J (assuming speedup of 1 second)

for 20 network transfers/hour (notifications, email, etc.), with 1 sec speedup
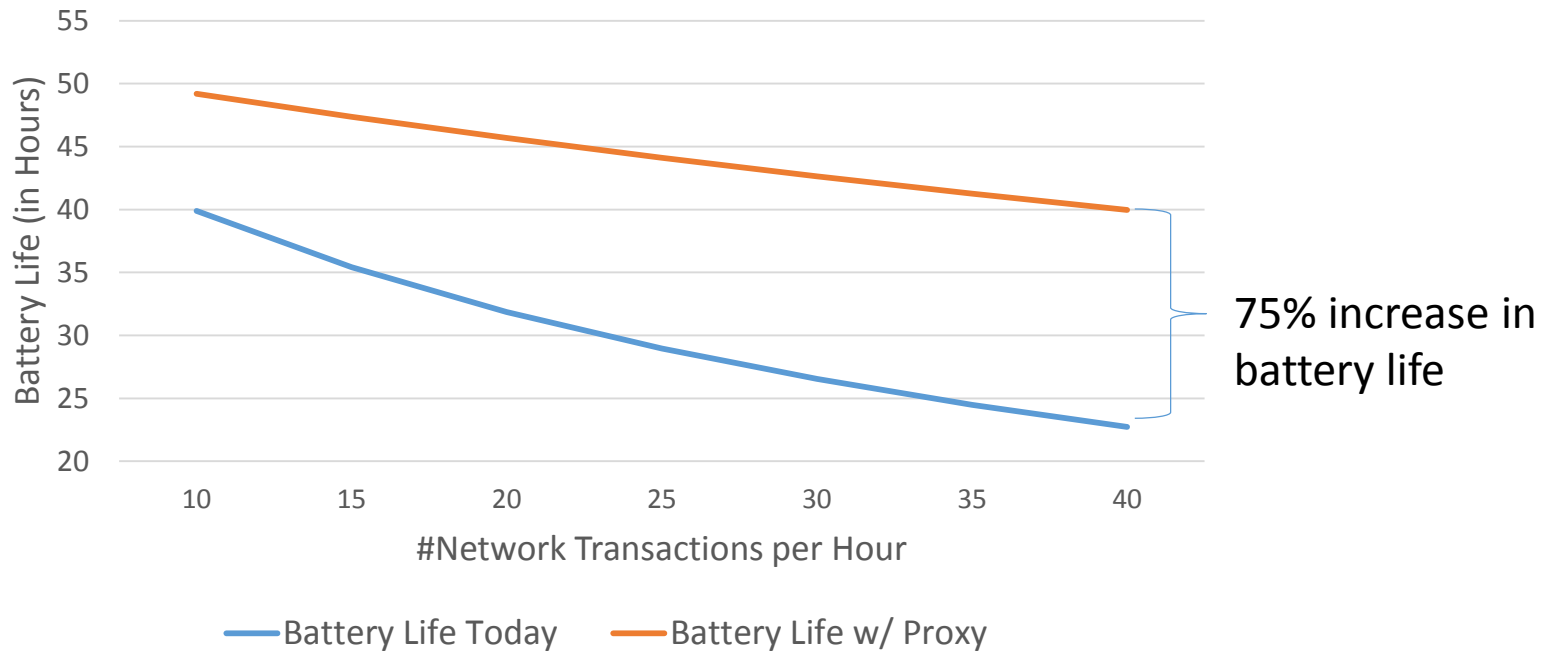total energy savings per 24 hr. = 6624 J
→ Saving of **26%** in a 1500 mAH cell phone battery*

* Samsung Standard LI-ION battery with rating of 1500mAh/3.7Vdc

especially good for mobile
battery life improvement

calculated for a 30 msec speedup / network transaction

75% increase in battery life

Y-axis: Battery Life (in Hours) — 20, 25, 30, 35, 40, 45, 50, 55
X-axis: #Network Transactions per Hour — 10, 15, 20, 25, 30, 35, 40

—— Battery Life Today   —— Battery Life w/ Proxy

these types of saving occur across the board for all battery types and all types of mobile devices

* Samsung Standard LI-ION battery with rating of 1500mAh/3.7Vdc

saving bandwidth….

security, traffic, tracking

locating objects of interest

customer queue analytics

current approach

- upload the captured video to the cloud for remote analysis

observations

- too much data captured per hour (>10GB/hour)

- bandwidth limits scale and use of system

- unable to support near real-time tracking & security

Microsoft

saving network bandwidth
(wireless video surveillance)

Faces detected in MSR lounge (III)

<10% frames capture objects of interest
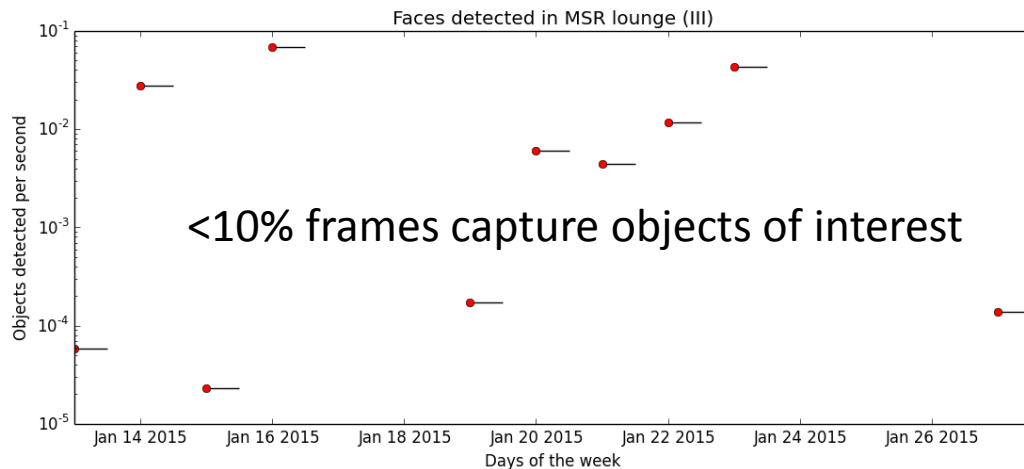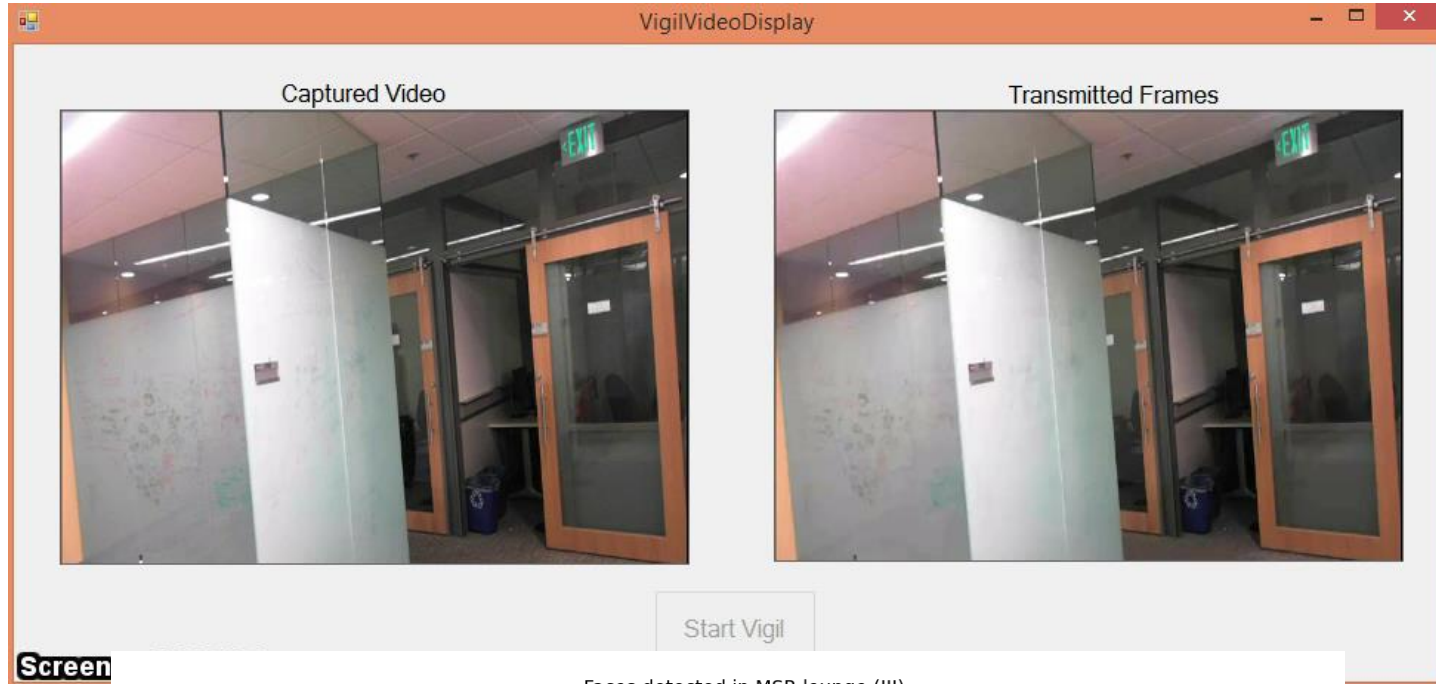
saving network bandwidth
(parking spot detector)
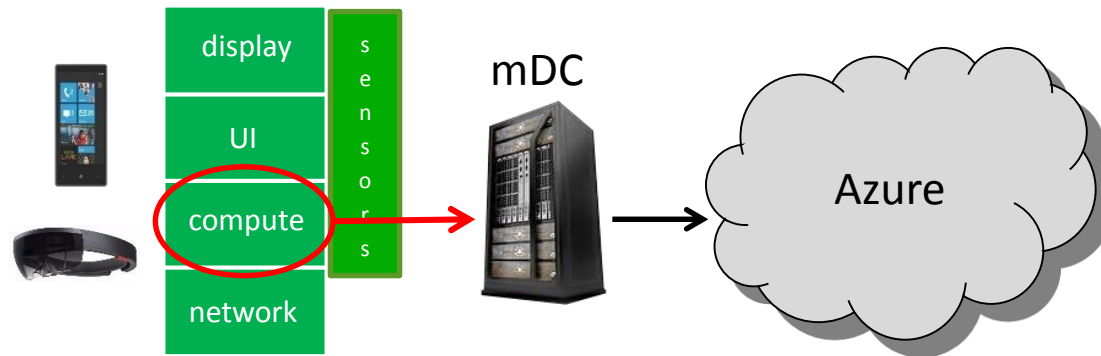
a couple of on-going problems

offloading computation

remote execution reduces energy consumption and improves performance

challenges
- what to offload?
- how to dynamically decide when to offload?
- how to do so with minimum programmer effort?
- how to support multi-tenancy with bullet-proof privacy?

programming frameworks for cloud offloading

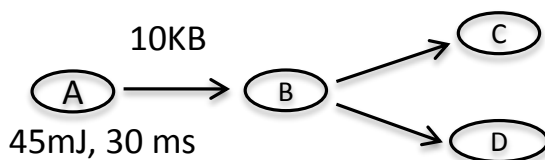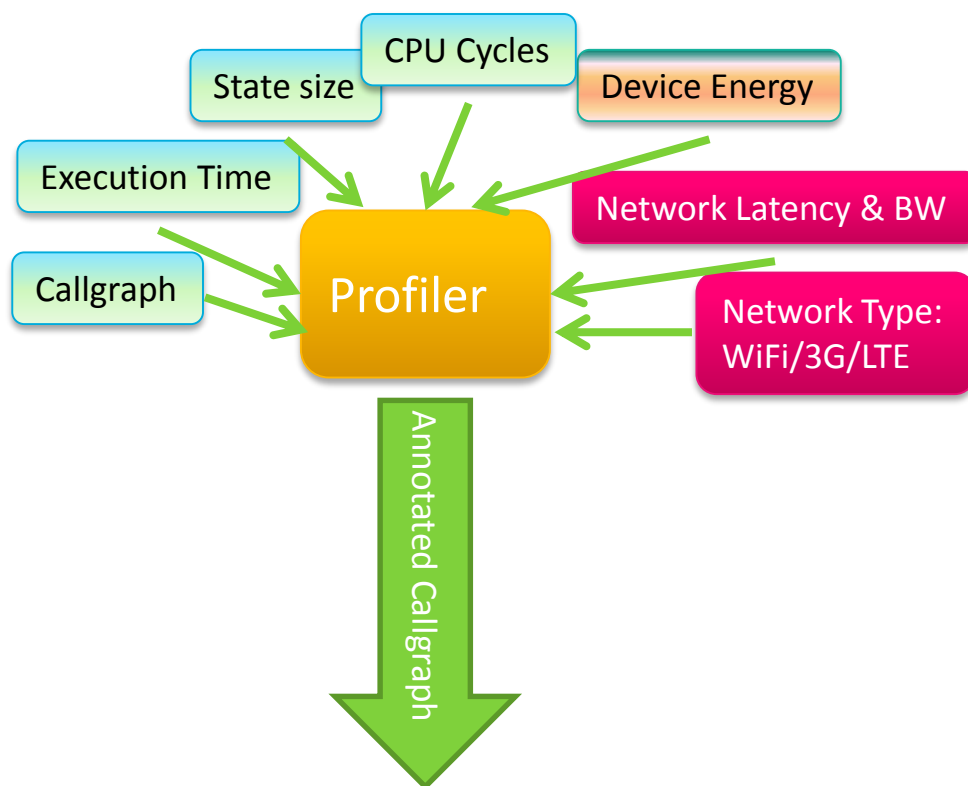|	Microsoft's MAUI	Intel's CloneCloud	USC's Odessa
remote execution unit	methods	threads	tasks

- MAUI exploits .NET framework to dynamically partitioning & offload method execution [MobiSys'10]

- CloneCloud supports existing applications, but requires tight synchronization between cloud and phone [EuroSys 2011]

- Odessa creates a data-flow graph to exploit parallelism [MobiSys 2011]

all have a profiler & a solver

also see: http://elijah.cs.cmu.edu/

Microsoft

MAUI's profiler and decision engine

profiler

CPU Cycles

State size

Device Energy

Execution Time

Network Latency & BW

Callgraph

Profiler

Network Type: WiFi/3G/LTE

Annotated Callgraph

10KB

A → B → C

B → D

45mJ, 30 ms

decision engine:

partition a running app – use Integer Linear Program

Example – Maximize:

$$\sum_{v \in V} (I_v \times E_v) \quad - \quad \sum_{(u,v) \in E} (|I_u - I_v| \times C_{u,v})$$

energy saved   cost of offload

Such that:

$$\sum_{v \in V} (I_v \times T_v) \quad + \quad \sum_{(u,v) \in E} (|I_u - I_v| \times B_{u,v}) <= Lat.$$
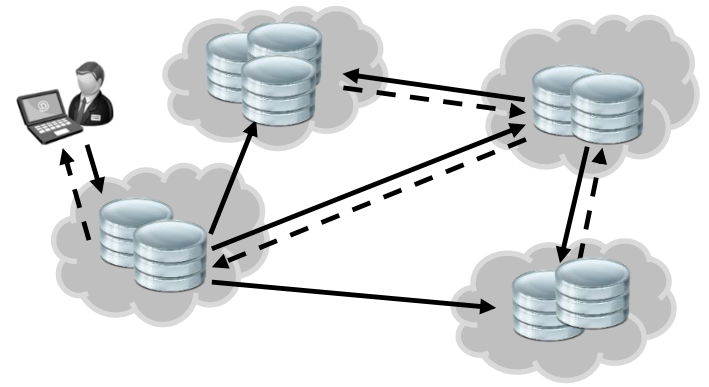
execution time   time to offload

and

$I_v <= R_v$ for all $v \in V$

Microsoft

geo-distributed analytics

lots of data being generated at the edges, need support for sophisticated analysis

possible solution(s)

- pull all data into a central data center; answer queries from there
- leave data where it is collected; fetch on demand per query

costly and wasteful; not realtime
very long latency; can't run Hive or Spark on WAN

Observations

- connectivity is expensive, low bw & high latency
- need to support near real-time triggers (e.g.. faults/ fire)
- some of the data is infrequently accessed

Microsoft

geo-distributed analytics

allow data & query tasks to be placed at any site
- some datasets remain at the edge; others move to resource-rich DCs
- make job schedulers' robust to high latency by pipelining

mimic *optimal* data & task placement
- minimize average query latency
  - E.g.., move data iff the cumulative *shuffle volume* of its queries exceeds data size
  - Eg., place network-heavy tasks on a site where there is more data to be read

Microsoft

recapping benefits of mDCs

latency reduction

- serve static content immediately
- SSL termination / split TCP
- edge to DC protocol enhancements

bandwidth saving

- compression
- procrastination
- edge analytics

service & internet monitoring

reliable connectivity
- overlay networking
- path diversity

battery saving

- computation offloads
- client proxying

high-end game streaming
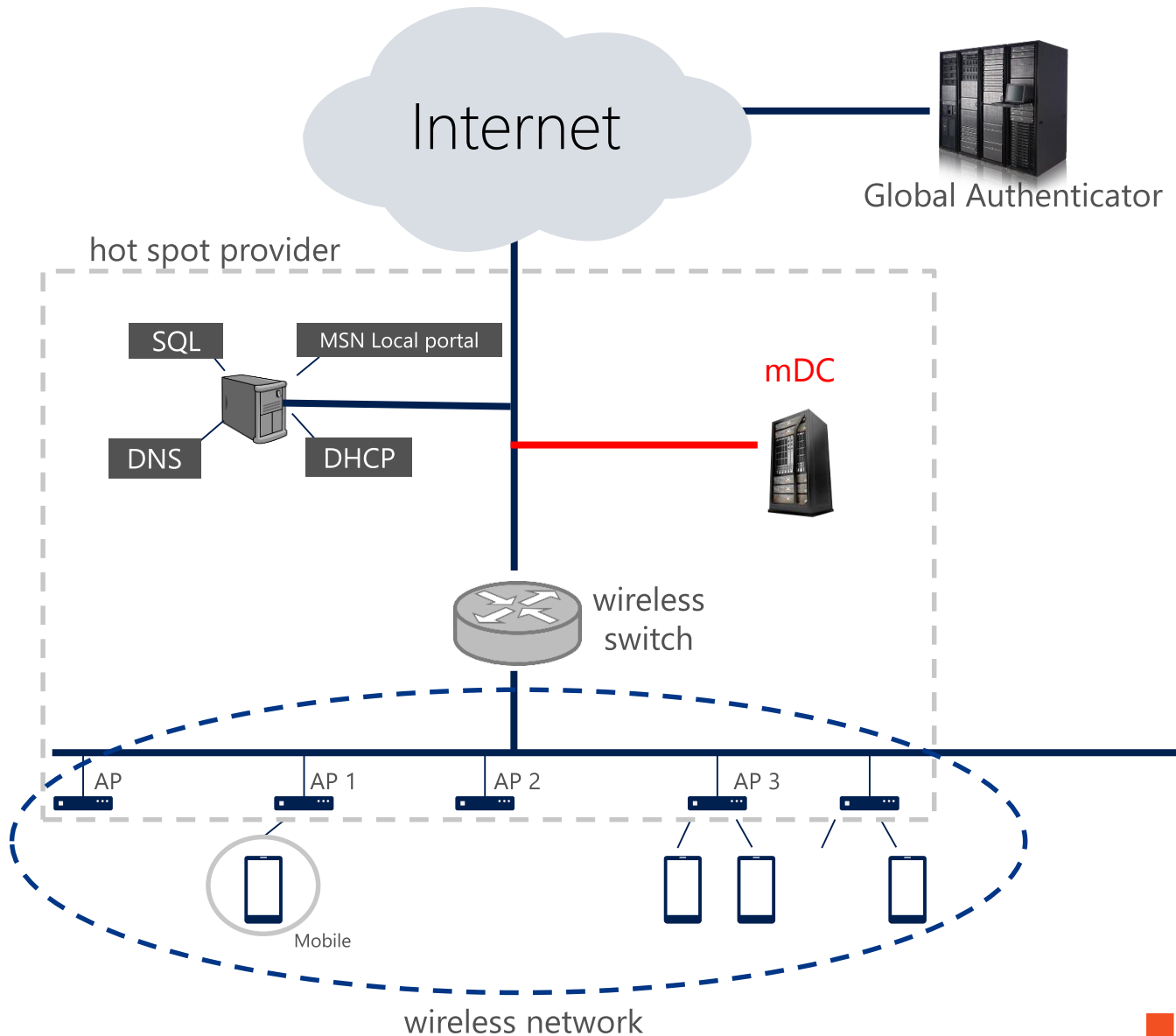- lower device cost
- reduce developer fragmentation

new services

protection against DoS
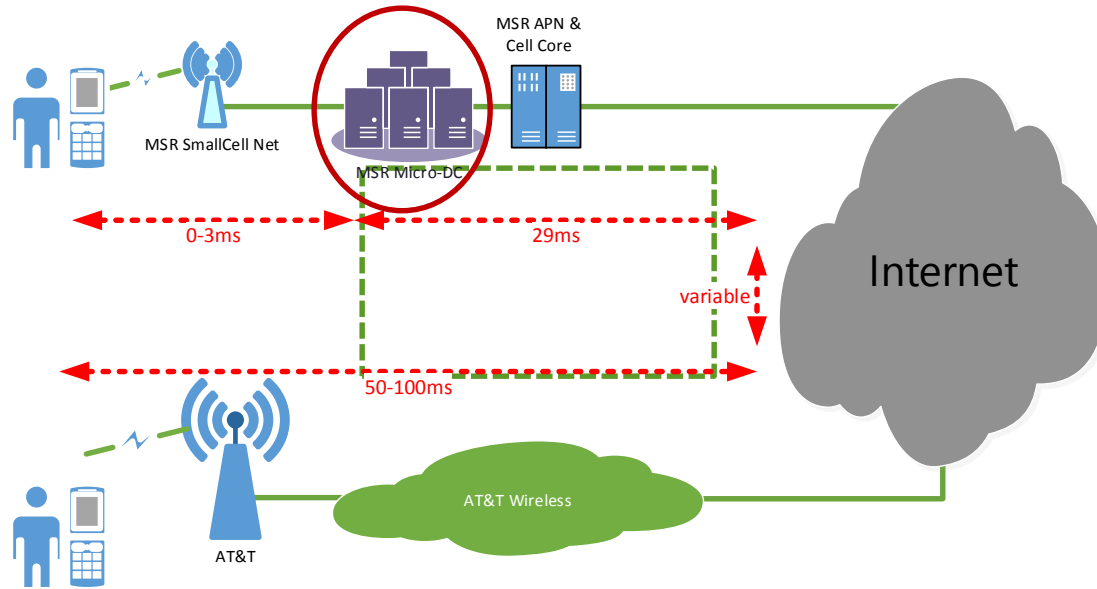
reduced load on DCs

deployment

mDCs with Wi-Fi or White-Fi

mDC with small cells

QCOM's Small Cell

MSR APN & Cell Core

MSR SmallCell Net

MSR Micro-DC

0-3ms

29ms

variable

50-100ms

Internet

AT&T Wireless

AT&T

Downlink	~110 Mbps
Uplink	~15 Mbps
RTT	~10 msec

```
Telnet 127.0.0.1

C:\>tracert any.edge.bing.com

Tracing route to any.edge.bing.com [204.79.197.200]
over a maximum of 30 hops:

  1    37 ms    34 ms    39 ms  172.26.241.113
  2     *       37 ms     *     172.26.236.2
  3    38 ms    38 ms    43 ms  172.26.96.11
  4    38 ms    39 ms    39 ms  172.26.96.193
  5    50 ms    41 ms    40 ms  172.18.3.241
  6    44 ms    37 ms    60 ms  12.249.2.25
  7    44 ms    43 ms    44 ms  12.83.180.6
  8    48 ms    47 ms    42 ms  12.83.180.14
  9    45 ms    52 ms    44 ms  cr81.st0wa.ip.att.net [12.122.5.197]
 10    93 ms   120 ms    43 ms  12.122.111.9
 11    45 ms    44 ms    46 ms  12.249.36.6
 12     *        *        *     Request timed out.
 13     *        *        *     Request timed out.
 14     *        *        *     Request timed out.
 15    50 ms    50 ms    50 ms  origin.any.bing.com [204.79.197.200]

Trace complete.

C:\>
```

tracert from AT&T LTE to *any.edge.bing.com* (15 hops)

```
Command Prompt

C:\Users\sagarwal>tracert any.edge.bing.com

Tracing route to any.edge.bing.com [204.79.197.200]
over a maximum of 30 hops:

  1     *        *        *     Request timed out.
  2    42 ms    27 ms    39 ms  131.107.151.1
  3    43 ms    98 ms    26 ms  ge-3-0-0--401.icar-sttlwa01-02.infra.pnw-gigapop.net [209.124.190.238]
  4    35 ms    27 ms    39 ms  ae1--706.iccr-sttlwa01-03.infra.pnw-gigapop.net [207.231.240.1]
  5    32 ms    28 ms    27 ms  microsoft-1-lo-jmb-706.sttlwa.pacificwave.net [207.231.240.7]
  6    30 ms    29 ms    27 ms  ae0-0.wst-96cbe-1a.ntwk.msn.net [204.152.140.105]
  7     *        *        *     Request timed out.
  8     *        *        *     Request timed out.
  9     *        *        *     Request timed out.
 10    43 ms    27 ms    38 ms  any.edge.bing.com [204.79.197.200]

Trace complete.

C:\Users\sagarwal>
```

tracert from SC to *any.edge.bing.com* (10 hops)

Microsoft

the wave is coming ...

Akamai
Cloudlets

Dynamically manipulate images in the Cloud for responsive web design

Fast and Easy Access Control for your Web Applications

(intel)

Increasing Mobile Operators' Value Proposition With Edge Computing

Turn bit pipes into smart pipes with an Intel® architecture-based server embedded into a Nokia Siemens Networks* base station
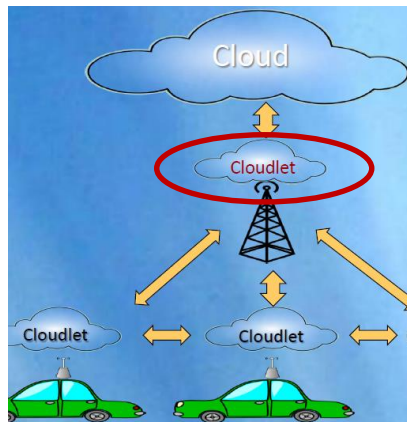
"local cloud are essential for backbone and core network scalability"
Dr. Geng Wu, Chief Scientist, **Intel** (Wireless World Research Forum, Vancouver, BC, Oct. 22, 2013)
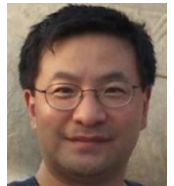
5G with **Undelay Networks** and **Local Cloud**

Cloud
Cloudlet
Cloudlet   Cloudlet

"cloudlets for reducing latency, security and reliability"

- Dr. David Soldani, VP **Huawei**
- (IEEE ICC, June 12, 2013)

Microsoft

...and it's becoming bigger
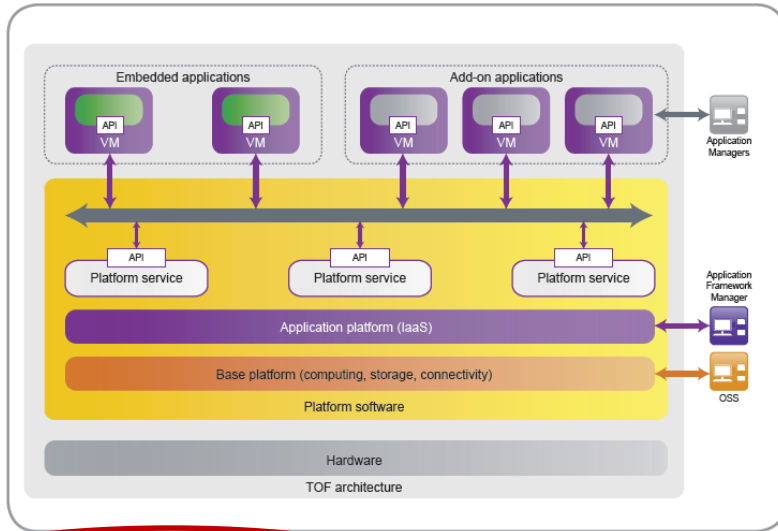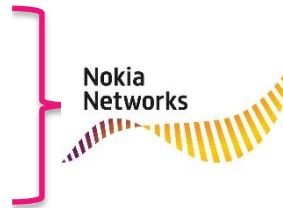## MOs moving towards edge services

Liquid Net

Figure 4: Base station application architecture

NETWORK WORLD

News

Nokia Siemens to merge cloud, base-station computing to boost performance

The company's Liquid Applications platform will use computing power in the cloud and in base stations, based on conditions

By Stephen Lawson, IDG News Service
February 24, 2013 04:06 PM ET

IDG News Service - Nokia Siemens Networks will expand the role of cellular base stations with a new platform that will store and deliver some application data locally, while tapping into information about subscribers and traffic to improve the process.

The company announced the system, called Liquid Applications, at an event in Barcelona on the eve of Mobile World Congress. Liquid Applications can improve consumers' mobile experience but cutting delays as well as delivering more relevant content, CEO Rajeev Suri

ETSI World Class Standards

Nokia Networks reveals ETSI mobile edge computing collaboration

IBM, Intel, Vodafone, and Huawei all on board

October 20, 2014 | By Michael Carroll

IBM and Nokia Siemens Networks Announce World's First Mobile Edge Computing Platform

Barcelona, Spain - 25 Feb 2013: Mobile World Congress: Nokia Siemens Networks and IBM (NYSE: IBM) announced today a collaboration to deliver the world's first mobile edge computing platform that can run applications directly within a mobile base station. This new platform allows mobile operators to create a truly unique mobile experience, relieve the ever increasing strain on network infrastructure and bring completely new solutions to market.

The new platform can accelerate the delivery of media-rich services by delivering content directly from the base station, ensuring enhanced quality of experience for consumers in the face of ever increasing data traffic growth. The platform also enables a new generation of

Microsoft

overheard at a recent conference (IEEE ICNC 2015)

"fog computing"

John Apostolopoulos
CTO &VP, Cisco, USA

Sujit Dey, Professor/ Director
Center for Wireless CommunicationsUCSD

Microsoft

it's hot in the research community as well...

there is plenty of research literature (incl. MSR's) that shows edge computing significantly enhances mobile experience

first paper →

Satya (CMU), Bahl (Microsoft), Caceres (AT&T), Davies (Lancaster)
The Case for VM-based Cloudlets in Mobile Computing
IEEE Pervasive Computing, October 2009

~ 900 citations

Cuervo (Duke), Balasubramanian (UMASS), Wolman, Saroiu, Chandra, Bahl (Microsoft)
MAUI: making smartphones last longer with code offload
ACM MobiSys conference, June 2010

~ 825 citations

Why a Cloudlet Beats the Cloud for Mobile Apps
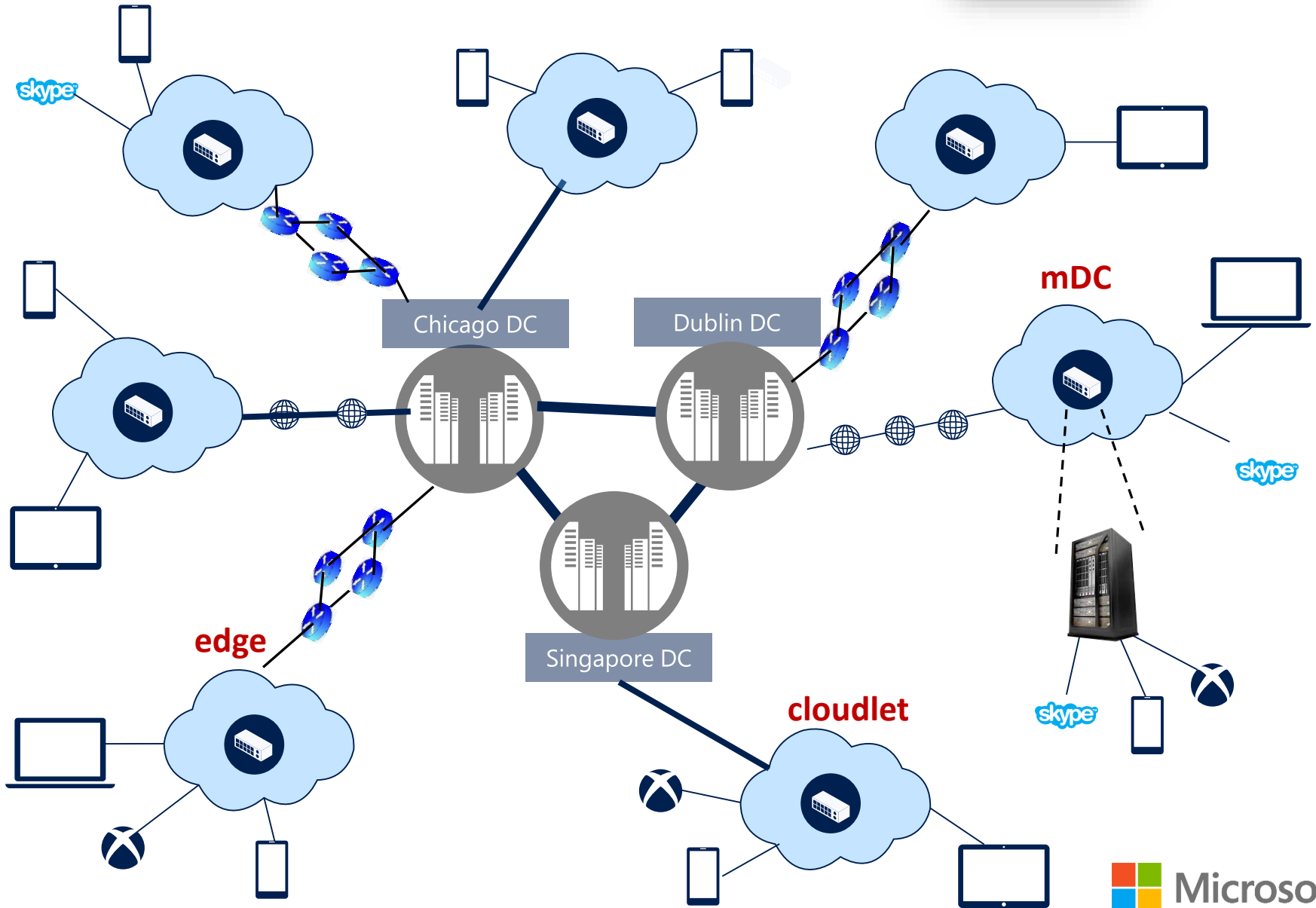
Posted on December 13, 2009 by lewisshepherd

Microsoft

cloud computing 2020

hyper scale data center

mDC

Chicago DC

Dublin DC

Singapore DC

edge

cloudlet

Microsoft

with mDCs (cloudlets) you can...

- develop new *(latency sensitive, CPU & battery intensive) (IoT)* applications, which *(dynamically)* partition themselves

- pursue infrastructure research in an emerging cloud platform, which promises to be pervasive

- deploy your own mDCs & connect them to Azure

Microsoft

merci!

mDC benefits -
app & game streaming

run any ecosystem's apps on resourced-starved devices by streaming them from the cloud

- circumvent client-side compatibility complexities

- with mDCs, reduce
  - latency -- keeping users engaged
  - jitter & packet loss – reduce user frustrating in highly interactive sessions
  - backbone bandwidth so both MOs and we pay less to other ISPs

note: standard proxy + split TCP insufficient for interactive traffic

Microsoft

mDCs can reduce dependency on cellular networks

offload to Wi-Fi aggressively

compress aggressively

already doing this

MobiSys 2014

procrastinate instead of prefetch
- many network apps. fetch data whether or not it is consumed
- **idea**: mDC fetches the data but holds on to it until user explicitly needs it
  - ✓ save cellular bandwidth <u>without</u> the latency penalty

Microsoft

procrastinate & save

few results on bandwidth saving

the system automatically decides what is not needed by the end-user

■ original app ■ what user needed ■ our savings

bytes (KB)

test applications

Microsoft
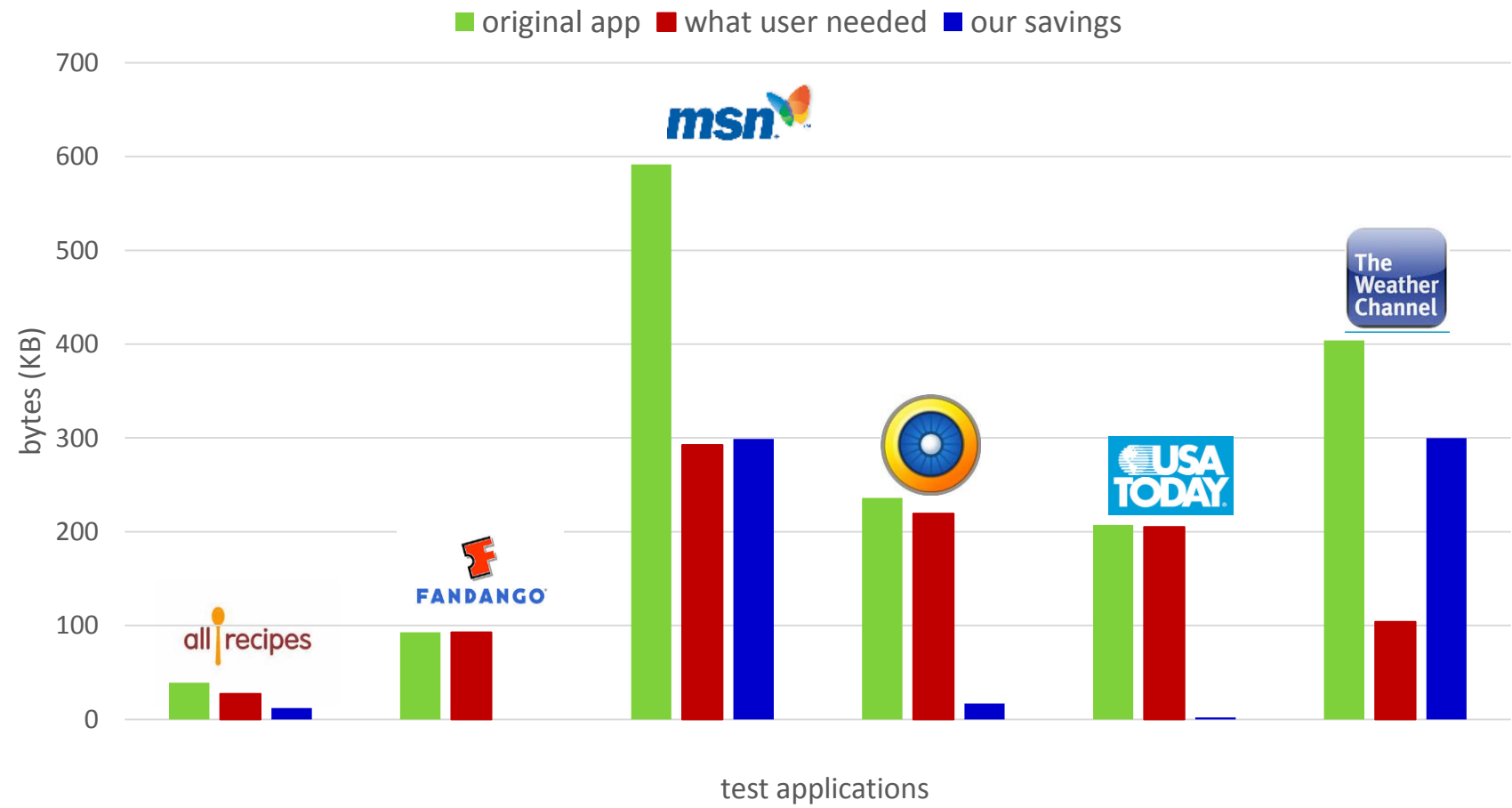
micro datacenter - benefits
reducing dependency on cellular networks (with procrastination)

get data only when needed (without mDC)

get data only when needed (with mDC)