Mega Trends Driving Architectures of Mobile Computing and IoT Devices

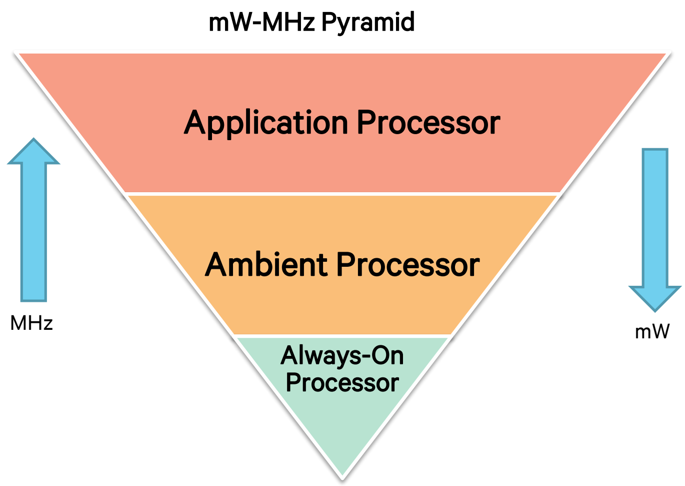
QUALCOMM

Karim Arabi
Vice President, R&D

Megatrends in ASICs and Semiconductors

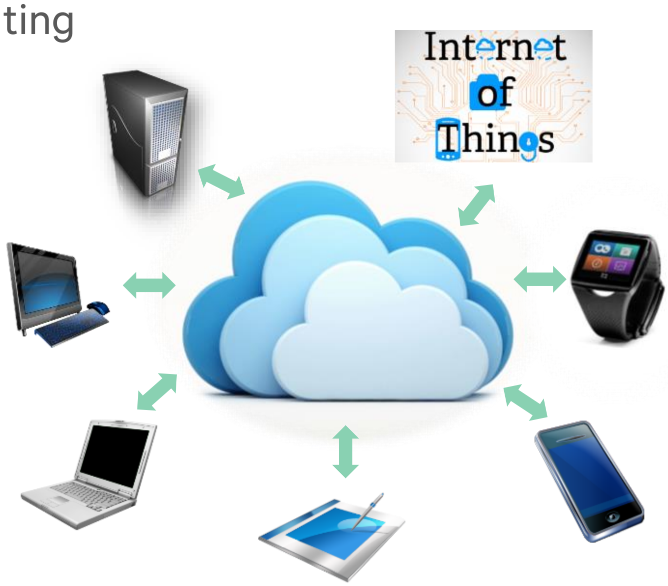
- Moore's Law Stall
- Always On (Always Aware)
- Cloud Computing / Big Data / Security
- Machine Learning / Deep Learning
- Propagation of Technologies Enabled by Mobile Computing
- Emerging Memories

Hierarchical Event Driven Computing



Cloud Computing

Big Data and abundant computing power are pushing computing to the **Cloud**



Instant Data generated by sensors and users are pushing computing to the **Edge**

Cloud Computing

- New Devices
- New Architectures
- High Performance Computing at Low Cost and Low Power
- Big Data Analytics
- Security

Machine Learning / Deep Learning

QUALCOMM

Machine Learning / Brain Inspired Computing

Human Brain	IBM Sequoia
~ 3.5 petabytes	1.6 petabytes
~ 20 petaFLOPS	16.3 petaFLOPS
~ 20 watts	7.9 mega watts

The Human Brain

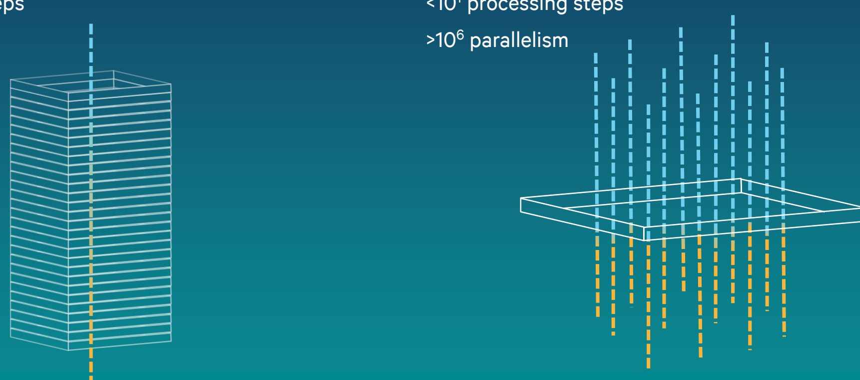
- is a massively parallel machine with ~86B neurons
- has no system clock, it is event driven
- has no hardware/software distinction
- performs processing and memory by the same components
- is a self-organizing, self healing system

 7

The Brain is a Massively Parallel Machine

$>10^6$ processing steps
$<10^1$ parallelism

$<10^1$ processing steps
$>10^6$ parallelism

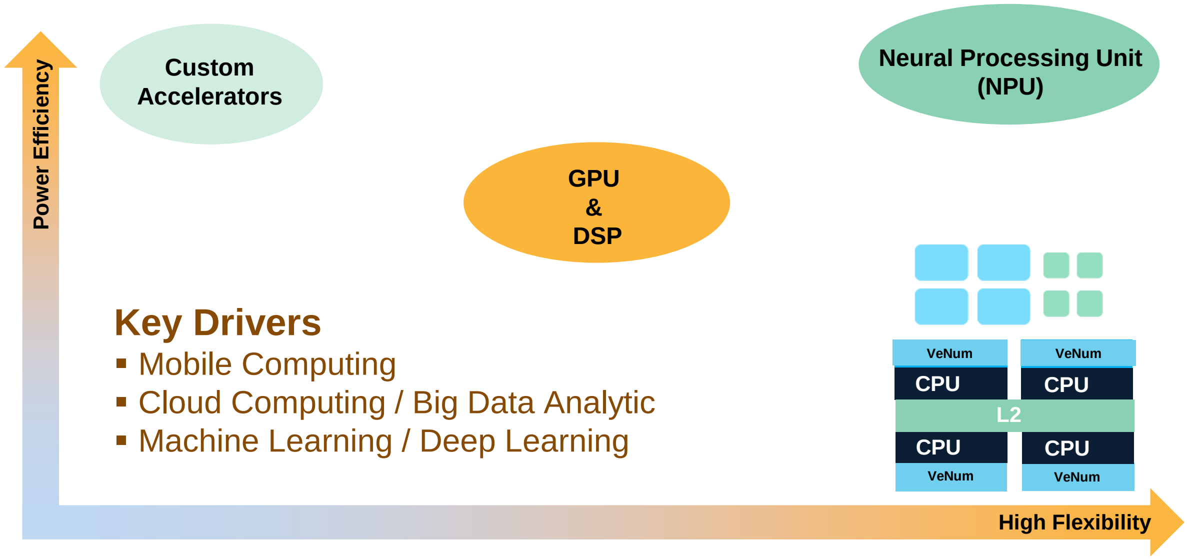


Modern computer
Dense, real-valued data

Human brain
Sparse 'Events" or "Spikes"

 8

Mobile Heterogeneous Compute Units to Lower Power

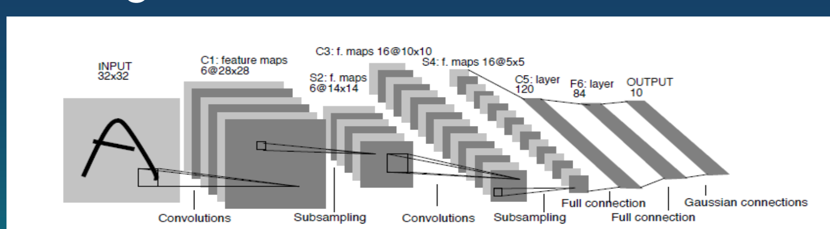


Key Drivers

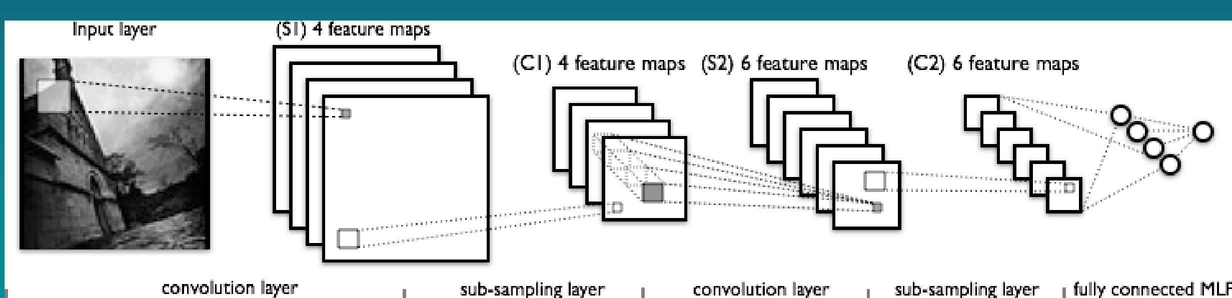
- Mobile Computing
- Cloud Computing / Big Data Analytic
- Machine Learning / Deep Learning

 9

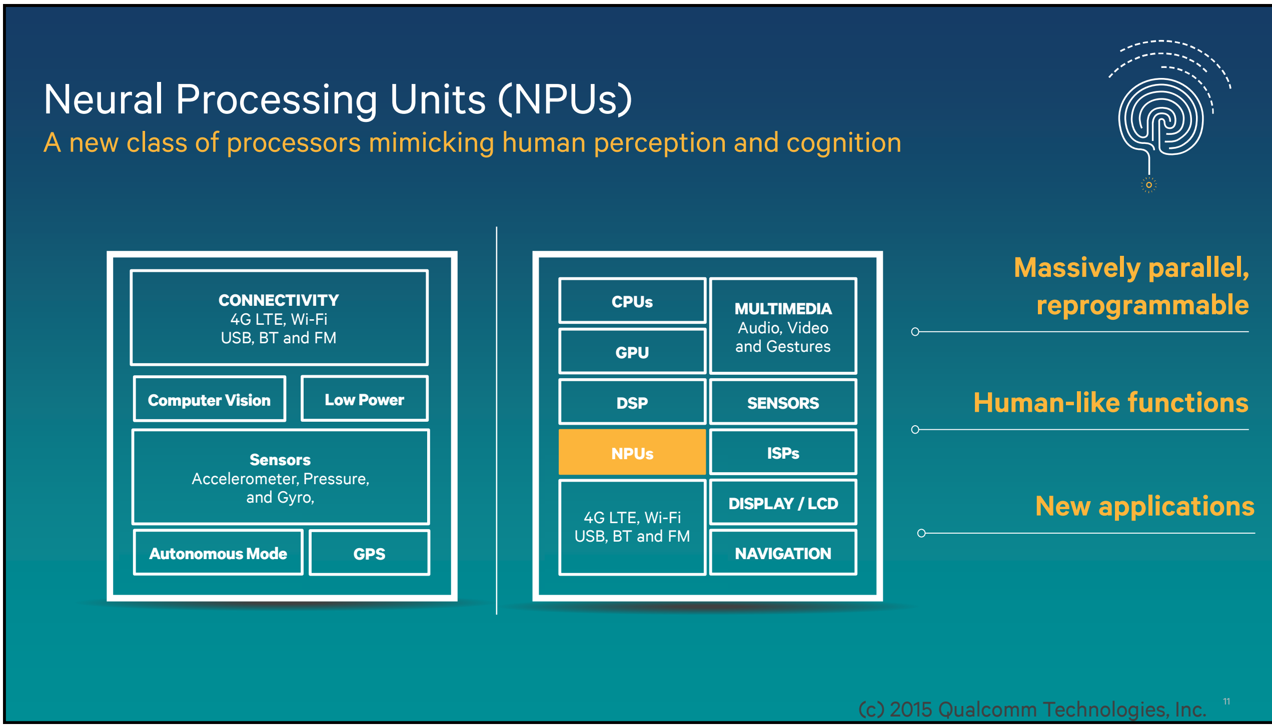
Deep Learning



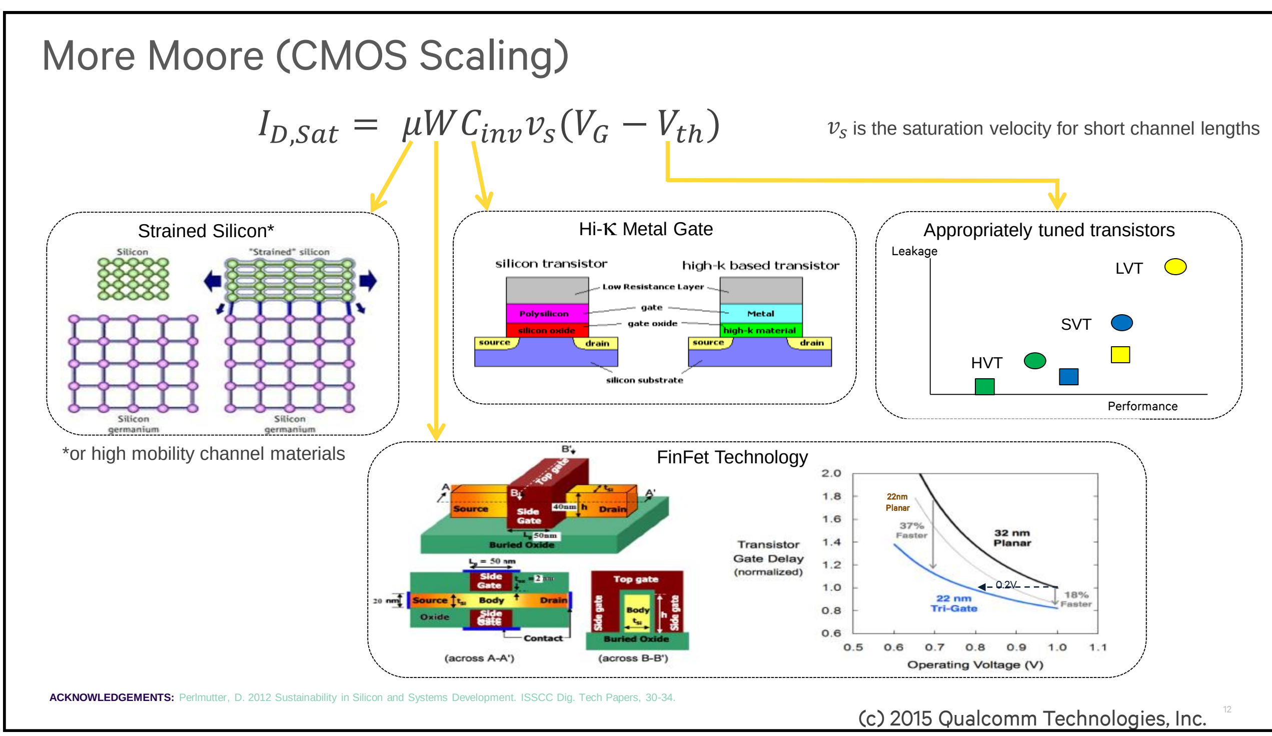
LeCun et al., 1998



 10



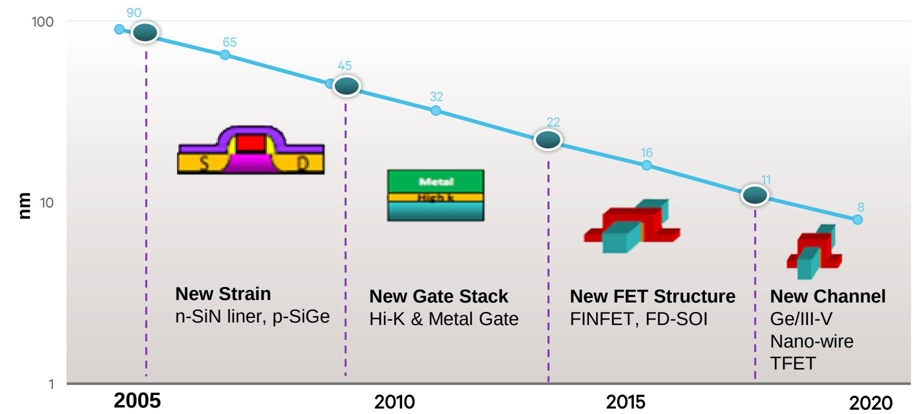
Neural Processing Units (NPUs)
A new class of processors mimicking human perception and cognition
CONNECTIVITY
4G LTE, Wi-Fi
USB, BT and FM
Computer Vision
Low Power
Sensors
Accelerometer, Pressure,
and Gyro,
Autonomous Mode
GPS
CPUs
GPU
MULTIMEDIA
Audio, Video
and Gestures
DSP
SENSORS
NPUs
ISPs
4G LTE, Wi-Fi
USB, BT and FM
DISPLAY / LCD
NAVIGATION
Massively parallel, reprogrammable
Human-like functions
New applications
(c) 2015 Qualcomm Technologies, Inc.
11



More Moore (CMOS Scaling)
$I_{D,Sat} = \mu W C_{inv} v_s (V_G - V_{th})$
v_s is the saturation velocity for short channel lengths
Strained Silicon*
Silicon
"Strained" silicon
Silicon germanium
Silicon germanium
*or high mobility channel materials
Hi-K Metal Gate
silicon transistor
high-k based transistor
Low Resistance Layer
Polysilicon
gate
Metal
silicon oxide
gate oxide
high-k material
source
drain
source
drain
silicon substrate
Appropriately tuned transistors
Leakage
LVT
SVT
HVT
Performance
FinFet Technology
Source
Side Gate
Top gate
Drain
Buried Oxide
Contact
Oxide
Body
(across A-A')
(across B-B')
Transistor Gate Delay (normalized)
22nm Planar
37% Faster
32 nm Planar
22 nm Tri-Gate
0.2V
18% Faster
Operating Voltage (V)
ACKNOWLEDGEMENTS: Perlmutter, D. 2012 Sustainability in Silicon and Systems Development. ISSCC Dig. Tech Papers, 30-34.
(c) 2015 Qualcomm Technologies, Inc.
12

QUALCOMM

More Moore (CMOS Scaling)



SOURCE: ITRS

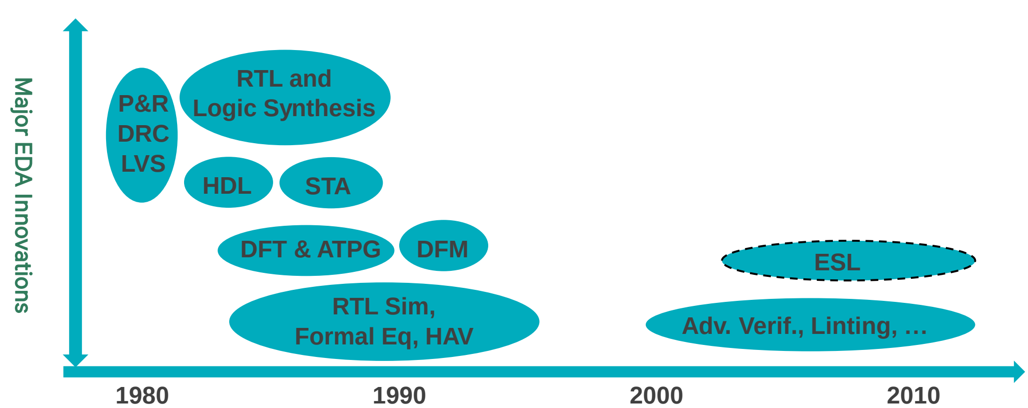
This Decade: **More Moore**
- Nano-wire
- TFET
- Ge/III-V

The 3rd Dimension
- Resistive memories
- MRAM
- 3DVLSI

 13

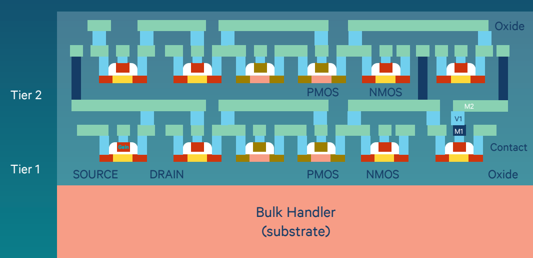
Rising Silicon Cost

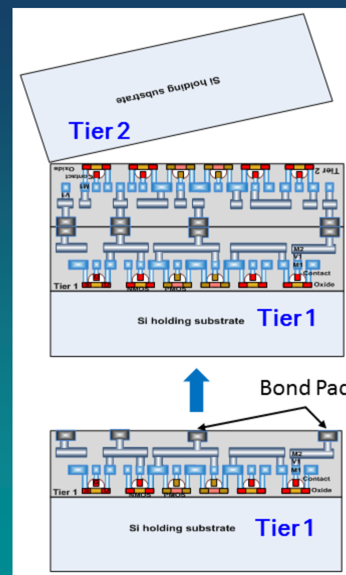
- Cost aspect of Moore Law is not scaling beyond 28nm
  - Expected to improve in 10nm but will continue to be a concern
  - EDA and architectural innovations are required



 14

More than Moore (3D-VLSI)





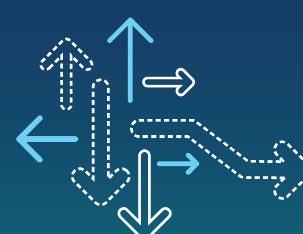
Key Attributes

- >100,000 inter-tier Vias/mm2
- 1 process node advantage

3D-VLSI PPA Gains

- 30% power saving
- 40% performance gain
- ~50% footprint reduction

15



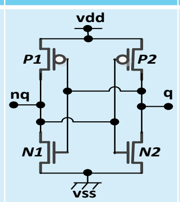
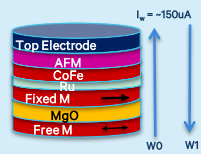
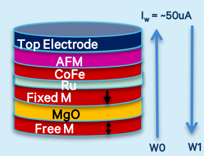
Emerging Memories

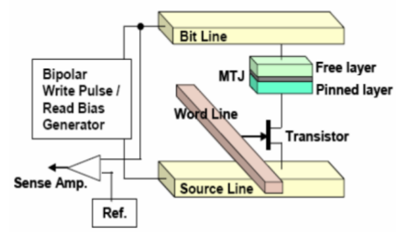
16

MEMORY

MRAM

- **MRAM power advantage**
  - **Zero retention/standby power from non-volatility, operating voltage much lower than SRAM**
  - **Reduction in write current also results in higher MRAM density (approaching 5x of SRAM density)**

	CMOS SRAM	In-plane MRAM	Perpendicular MRAM
Memory Element			
Magnetic Anisotropy	N/A	In-plane	Perpendicular
Switching Mechanism	Charge Racing	Spin-Transfer Torque (STT)	Spin-Transfer Torque (STT)
Energy per Cell per Write	0.52 pJ (40nm UMC)	W0: ~2x SRAM W1: ~1x SRAM	W0: ~24% SRAM W1: ~12% SRAM



MRAM Bitcell and R/W Circuit

 17

Emerging Memories

- MRAM, RRAM, …
- New memory architectures in mobile computing and data centers
- New computing architecture
- In memory compute

 18

QUALCOMM

Propagation of Mobile Computing Wave

19

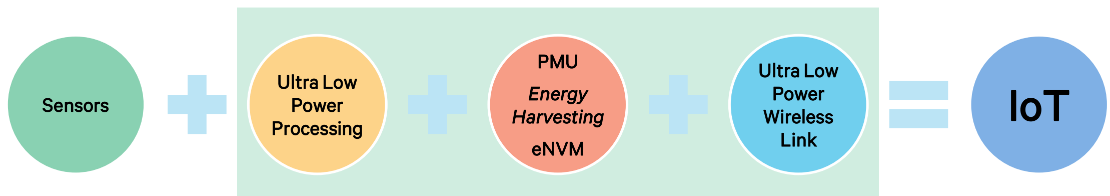
Propagation of Mobile Computing Wave to Other Markets

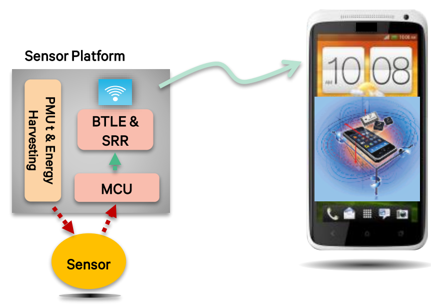
- IoT
- Wearables
- Drones and Robotics
- Automotive: Smart Cars
- Health

20

QUALCOMM®

Wearable and IoT Drive New Low Power Design Requirements





Conclusion

- Mobile computing enabled a new set of capabilities and applications
  - Heterogeneous computing, low power, low cost sensors
  - Always on, always aware
  - Cloud computing, Wearables, IoT
- Current Mega Trends impacting semiconductors growth and computing architecture
  - Propagation of technologies enabled by Mobile Computing
  - Always on, always aware
  - Cloud Computing /Big Data / Security
  - Machine Learning / Deep Learning
  - Emerging Memories
- New EDA innovations required


Thank you!
QUALCOMM