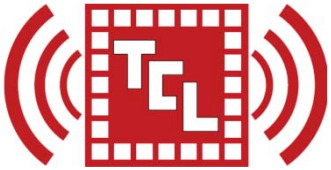
From Marconi to Moore

Circuits and Systems for Communications – Still a Challenge?

Andreas Burg



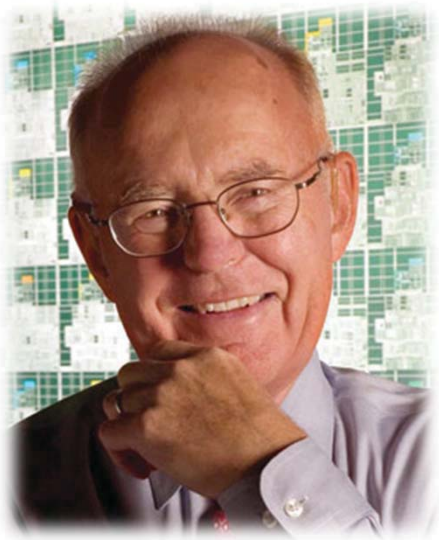
Acknowledgements: Prof. H. Meyr, M. Witte,
F. Borlenghi (RWTH-Aachen)

Visions that Changed the World by Marconi and Moore

Guglielmo Marconi

- *1897 : Wireless Telegraph Company*
- *1909 : Nobel Price in Physics*

„It is dangerous to put limits on wireless“



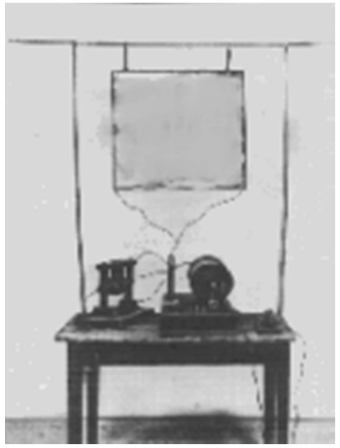
Source: Intel Corporation

Gordon Moore

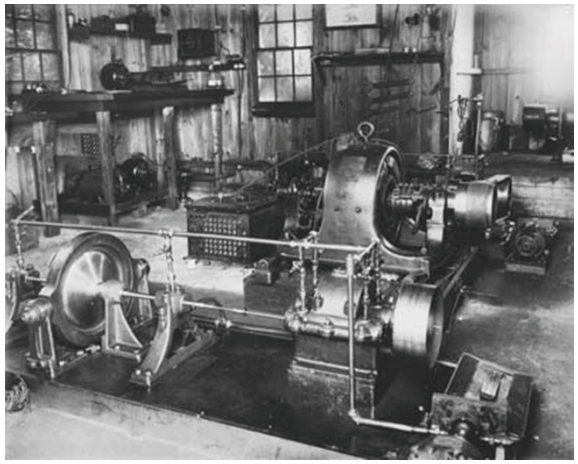
- 1968 : Co-founded INTEL Corporation
- 2005 : Marconi Society Lifetime Achievement Award

Moore‘s law (1965/75) **paces the evolution of integrated circuits until today**

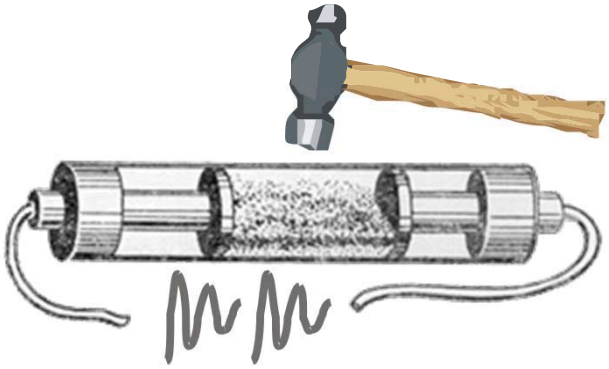
The Early Days of Wireless



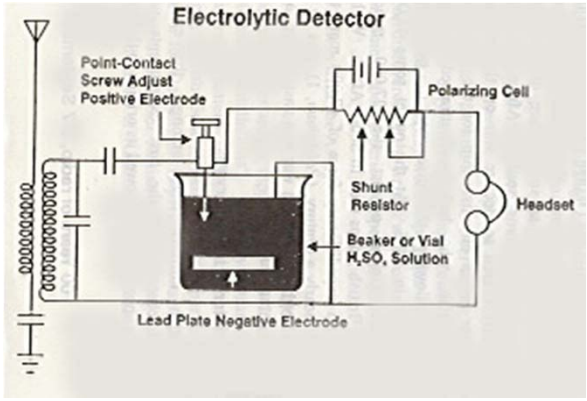
Marconi's first wireless telegraph (1895)



Mechanical continuous wave RF signal generator

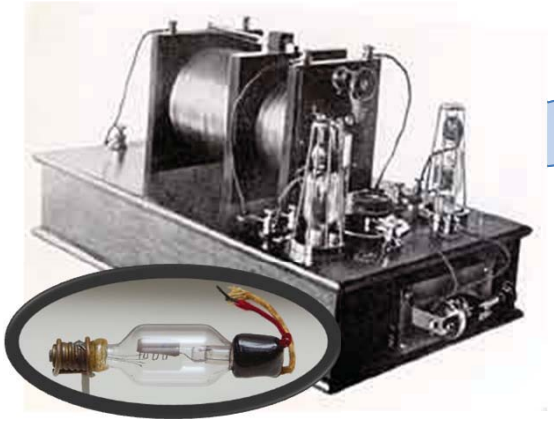


Core of an early days radio telegraph receiver



Electrochemical demodulator for voice ~1900

Microelectronics Enables Integration

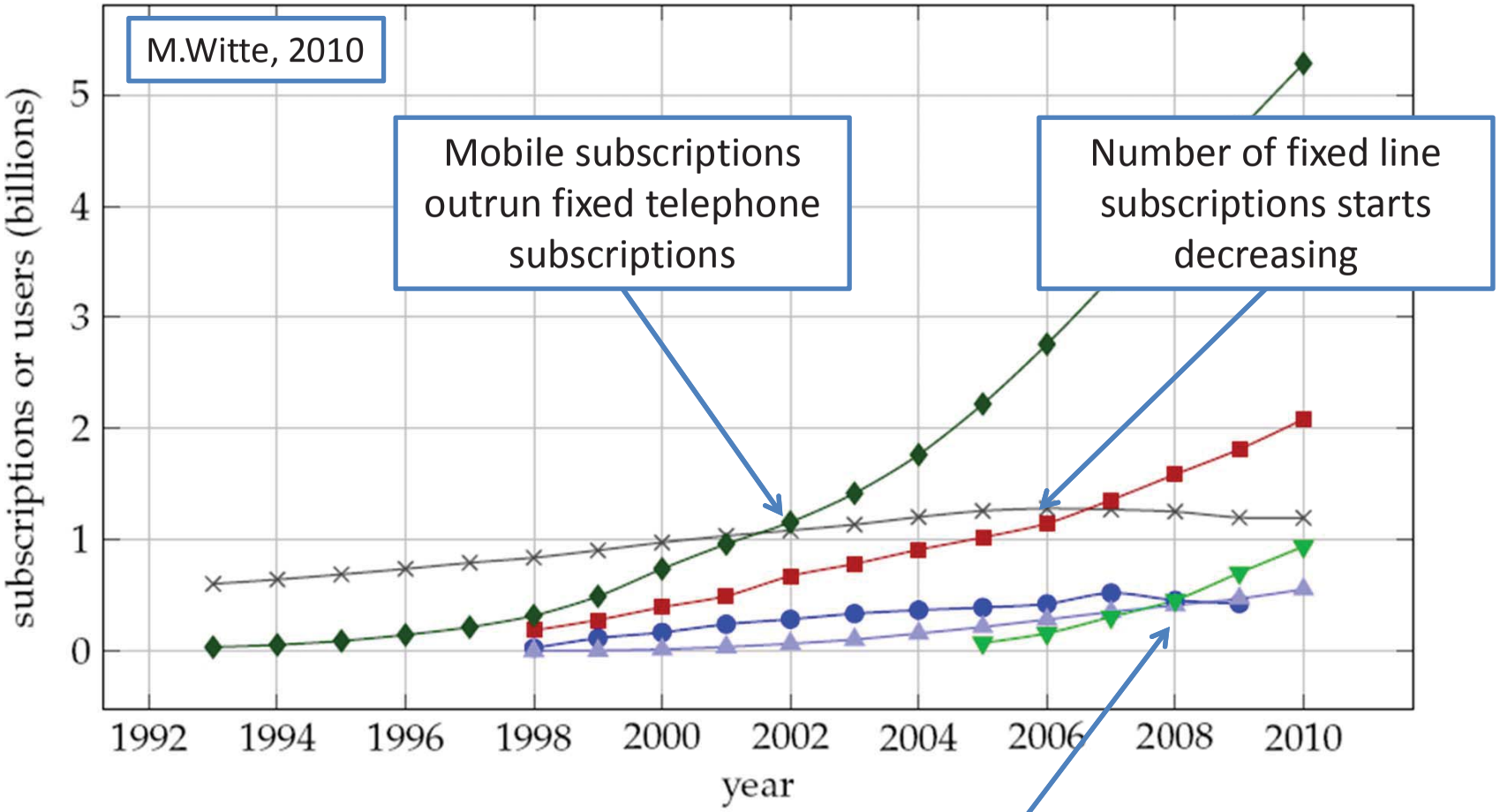


De Forest Audion radio receiver from 1906

Texas Instruments first silicon transistor (1954)

First transistor radio: TI Regency TR-1 (1954)

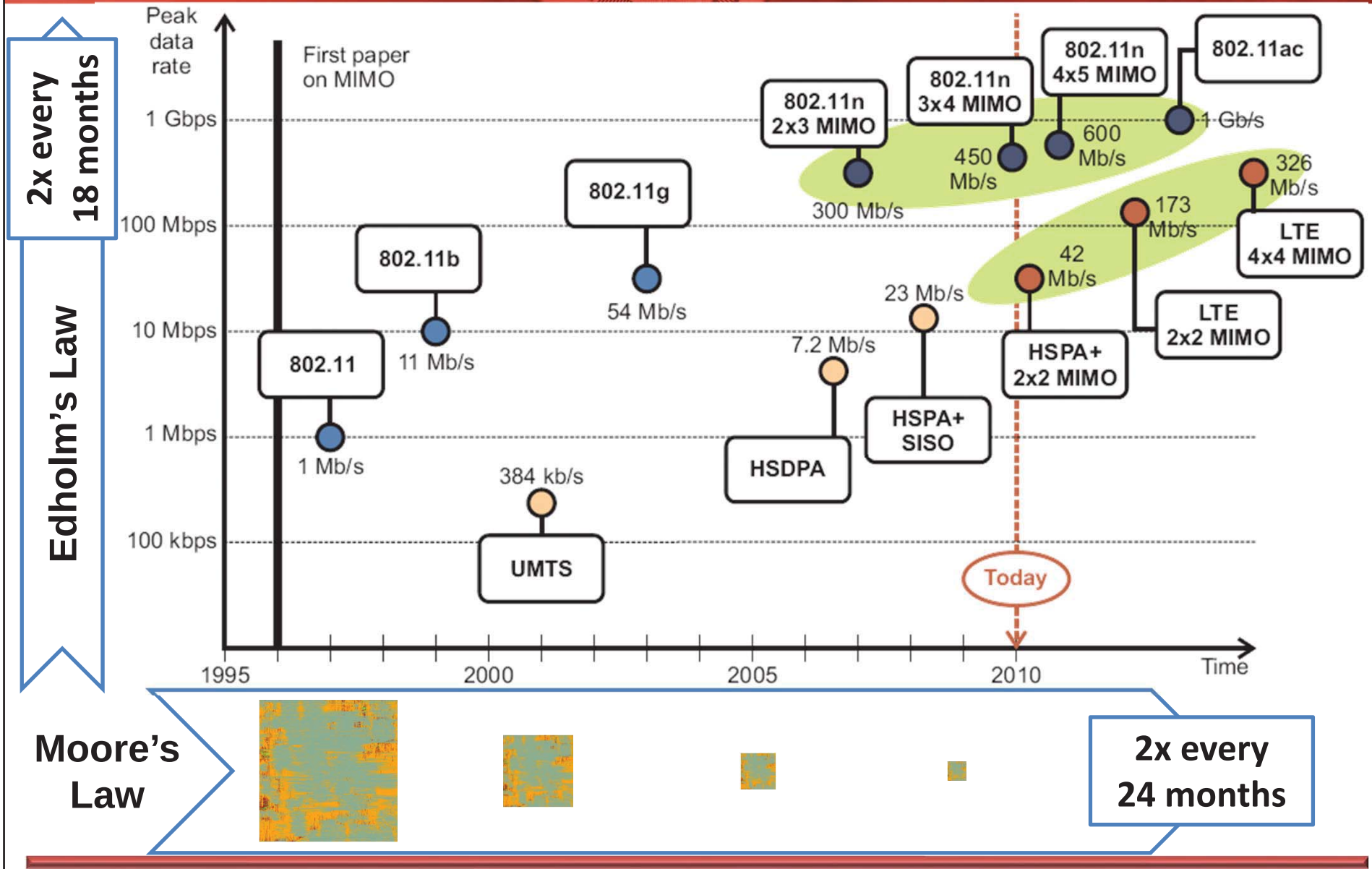
Evolution Toward Wireless Communications



Outline

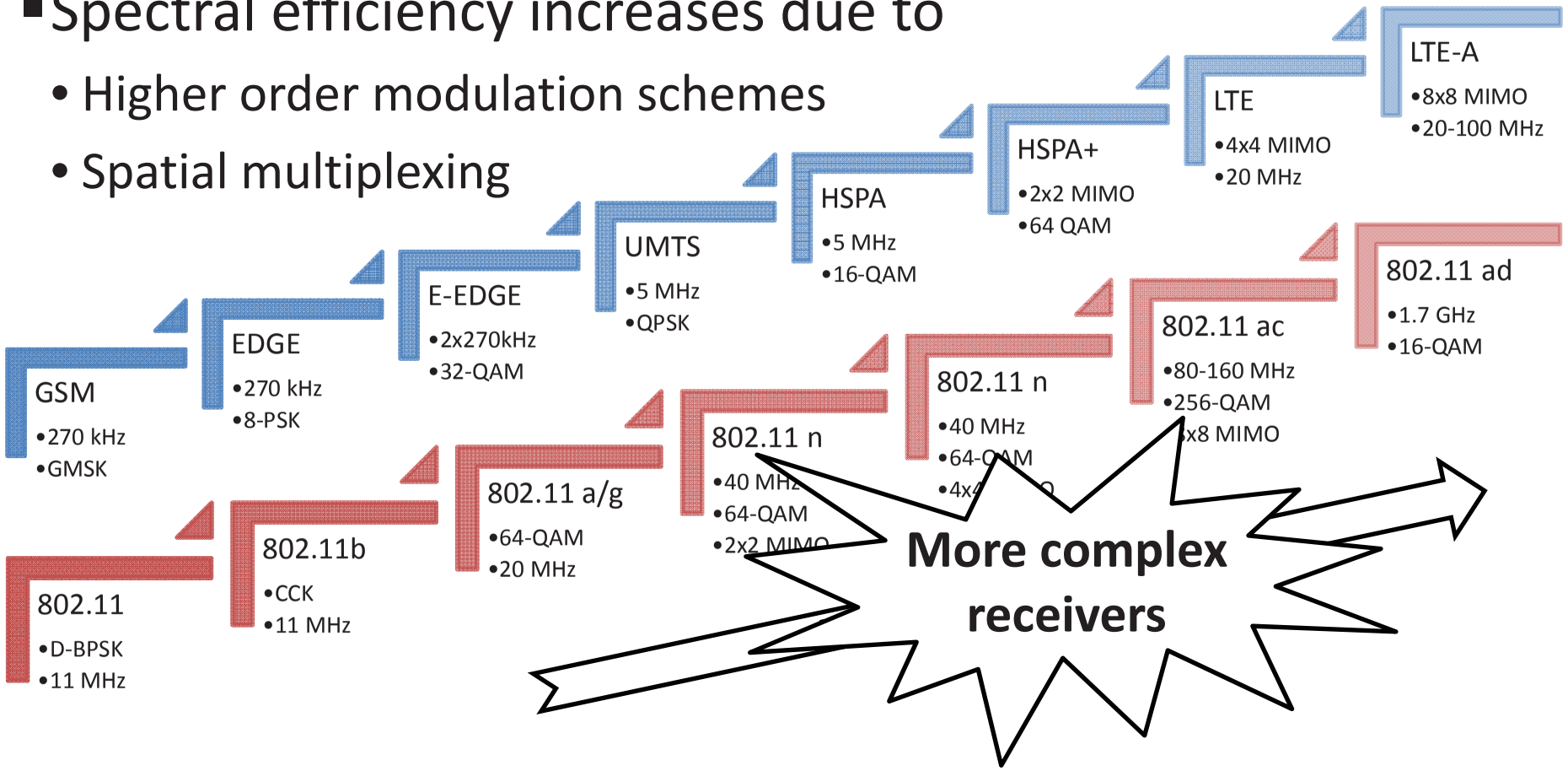
- Introduction: Some History
- Scaling Laws, Trends, and Observations
- Are there still challenges??
  - Some examples why I think YES
- The End of Moore's Law
  - The limit of wireless OR motivation for some more fancy research

Scaling Laws



Measures that Enable Higher Throughput

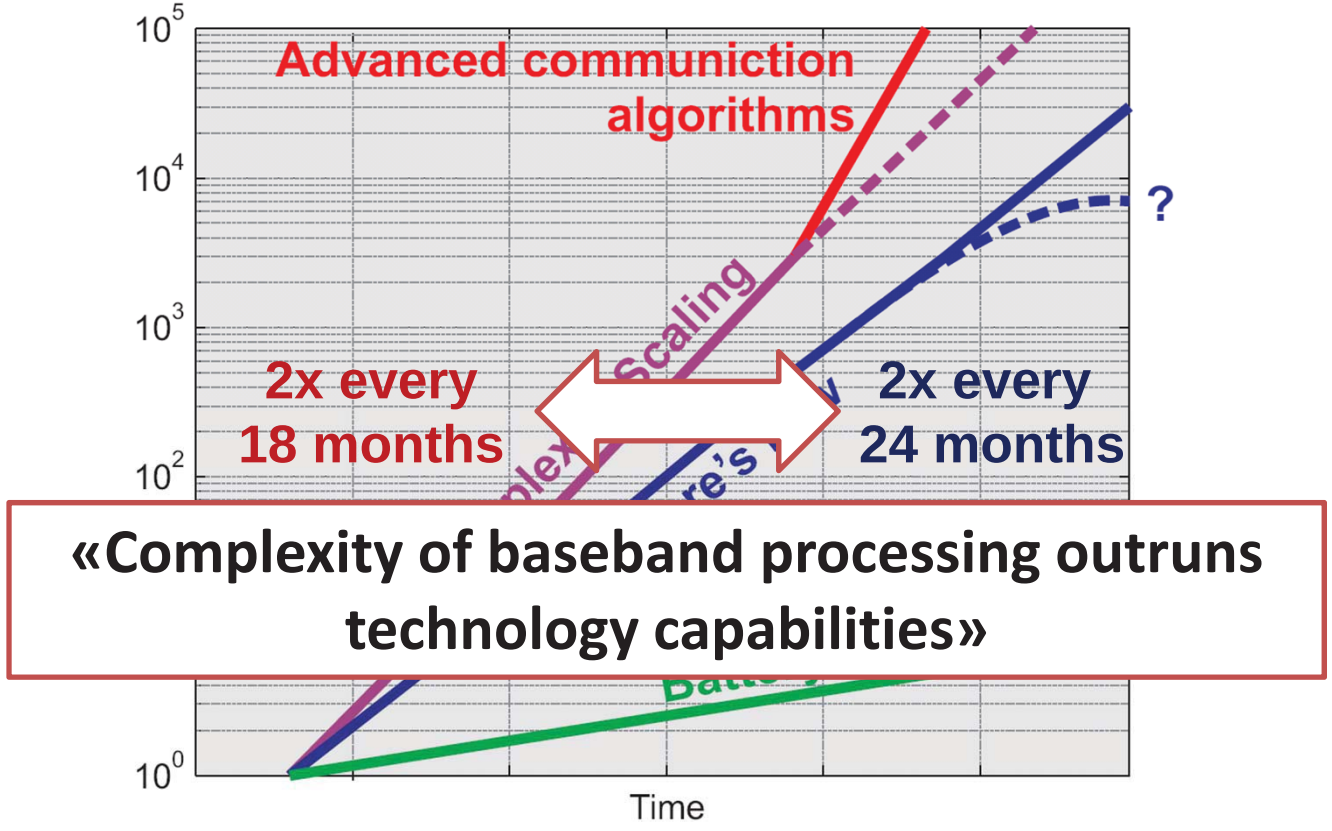
- New multiplexing schemes allow to allocate more bandwidth to a single user for higher peak throughput
- Spectral efficiency increases due to
  - Higher order modulation schemes
  - Spatial multiplexing



The Myth of Complexity…

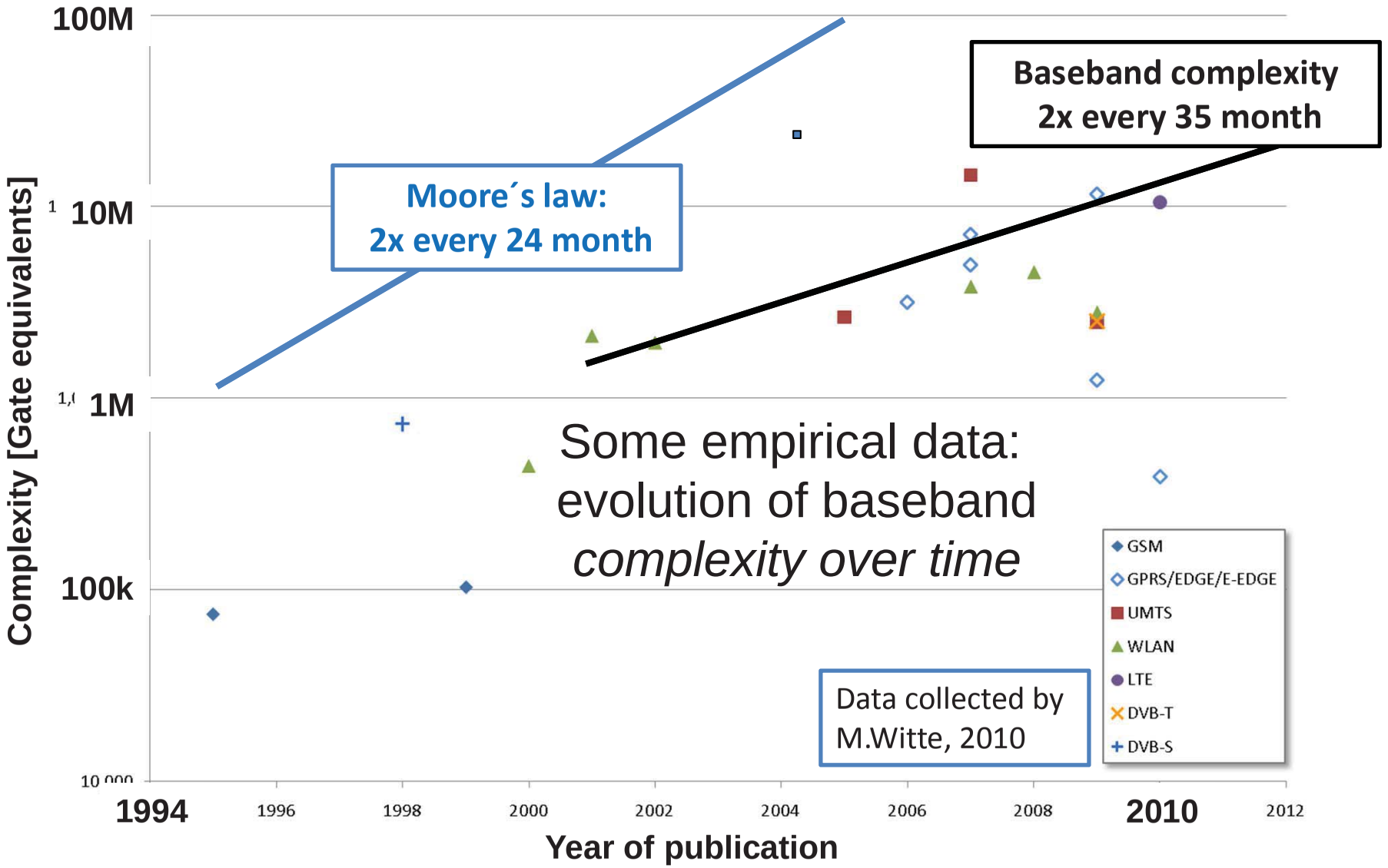
Imbalance between complexity and integration density

- Data rate doubles every 18 months
- Algorithm complexity grows (spectral efficiency)

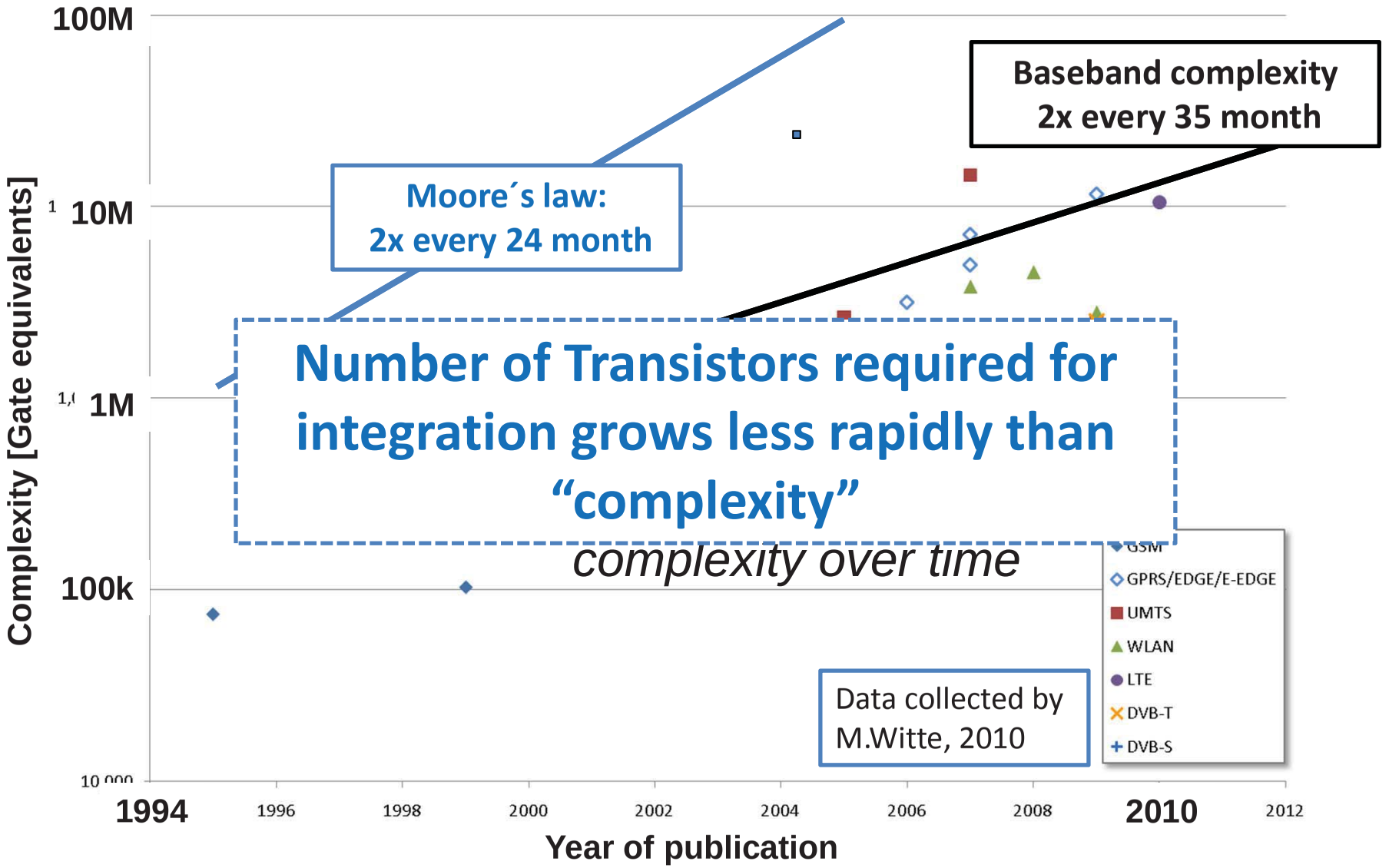


«Complexity of baseband processing outruns technology capabilities»

… Lets Check!

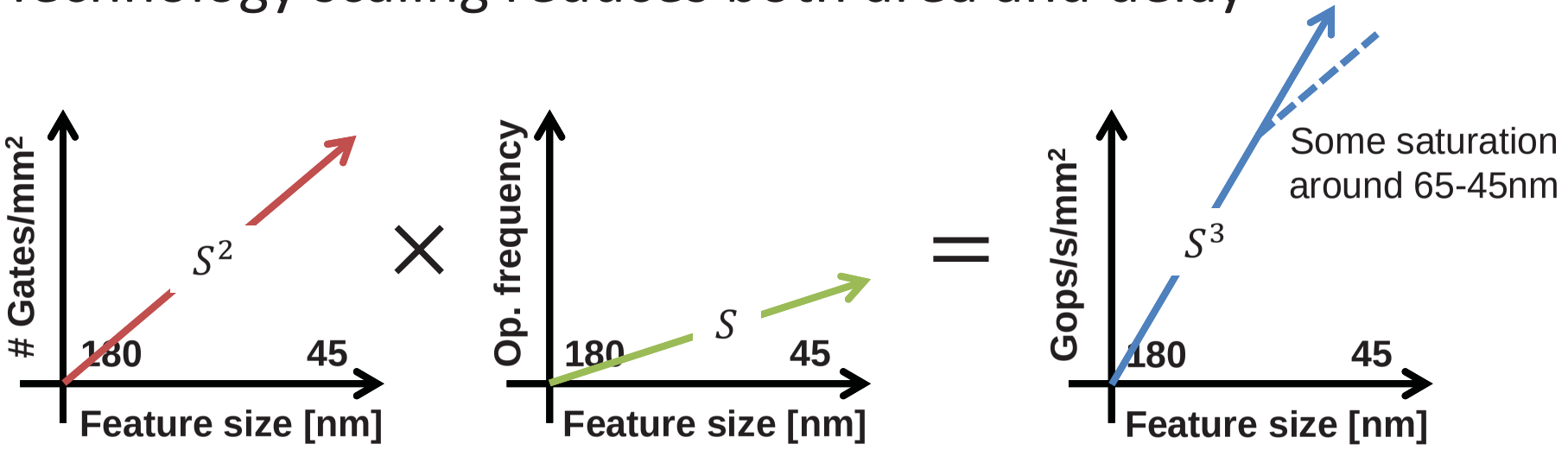

100M
10M
1M
100k
10 000
Complexity [Gate equivalents]
Moore´s law:
2x every 24 month
Baseband complexity
2x every 35 month
Some empirical data:
evolution of baseband
complexity over time
Data collected by
M.Witte, 2010
GSM
GPRS/EDGE/E-EDGE
UMTS
WLAN
LTE
DVB-T
DVB-S
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
Year of publication

… Lets Check!


100M
10M
1M
100k
10 000
Complexity [Gate equivalents]
Baseband complexity
2x every 35 month
Moore´s law:
2x every 24 month
Number of Transistors required for integration grows less rapidly than "complexity"
complexity over time
GSM
GPRS/EDGE/E-EDGE
UMTS
WLAN
LTE
DVB-T
DVB-S
Data collected by
M.Witte, 2010
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
Year of publication

GOps/s/mm² Increases Faster than Transistor Density

- Technology scaling reduces both area and delay

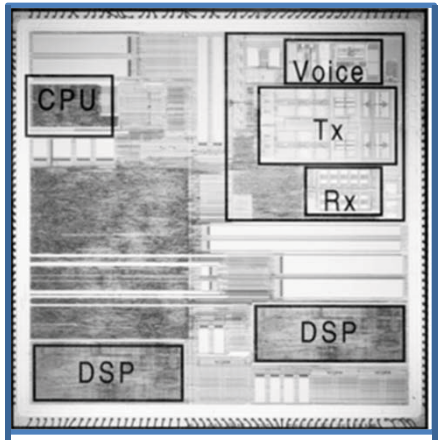


Example

	180nm	90nm	45nm
Clock freq.	100 MHz	200 MHz	400 MHz
16x16 Mult + overhead (50/50)	20k um²	5k um²	1.25k um²
	5 Gops/s/mm²	**40 Gops/s/mm²**	**320 Gops/s/mm²**

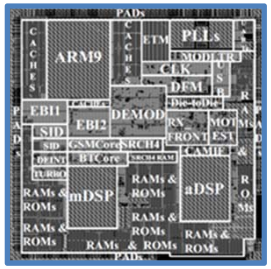
An Interesting Observation: Chips Do Not Shrink…

Some examples of digital cellular ASICs from ISSCC



2000

- Liu et al.
- 112mm2
- 250nm

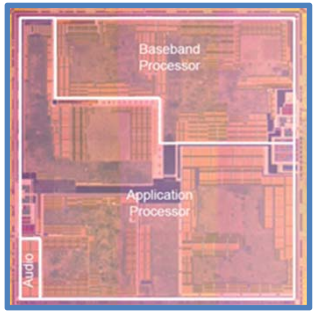


2004

- Uvieghara et al.
- 46mm2
- 130nm

2006

- Luftner et al.
- 43mm2
- 90nm



2009

- Shirasaki et al.
- 66mm2
- 45nm

2G → 2.5G → 3G → 3.5G

… But they Integrate More Modem Functionality

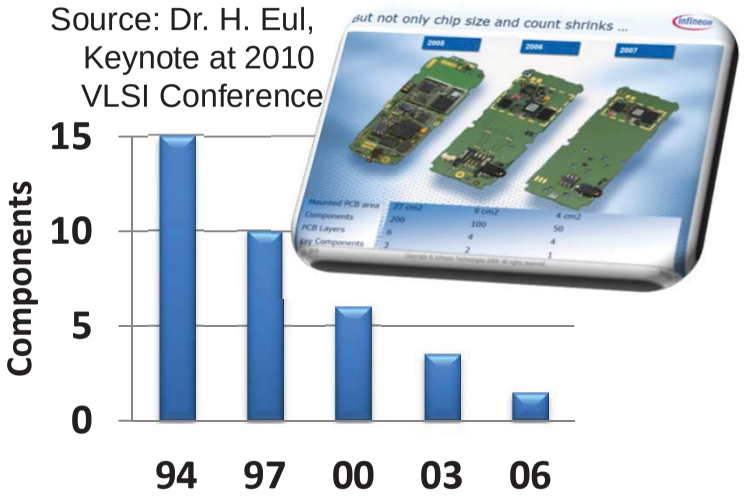
- Cellular modems require multi-standard support
- Nevertheless the number of discrete modem components decreases rapidly

GSM
EDGE
WCDMA
HSPA+
LTE
LTE-A

Legacy support guarantees coverage

2G (GSM/EDGE)
3G (WCDMA/HSPA)

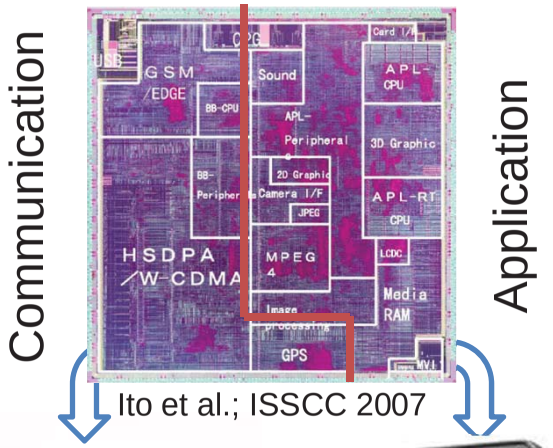
Source: Dr. H. Eul, Keynote at 2010 VLSI Conference



- Reduces cost (PCB, packaging, and manufacturing
- More space for battery and display

… and Many Other Functions in a Single Chip

- Integration of application and modem functionality
  - Additional air interfaces, connectivity options, and storage
  - 3D Graphics and Video
  - Powerful application processors



Ito et al.; ISSCC 2007

3GPP Modem covers 40% of the chip area

What drives us towards better technologies

Power consumption and energy efficiency

Standby/voice

Power consumption

$P[mW]$

- Determined by leakage and standby activity

Data download

Energy efficiency

$E\ [nJ/bit]$

- Determined by active power consumption

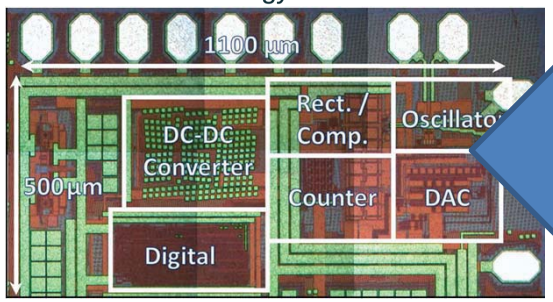
What drives us towards better technologies

Power consumption and energy efficiency

Standby/voice

Data download

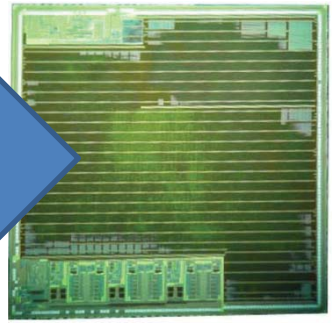
Technology: 0.18 um



Simple OOK radio for sensor nodes
0.18 nJ/bit (complete transceiver)

J. Ayers, *et al..,"An Ultralow-Power Receiver for Wireless Sensor Networks," JSSC 2010*

Technology: 0.18 um



802.11n WLAN transceiver
3 nJ/bit (digital PHY/MAC only)

P. Petrus, *et al., "*An Integrated Draft 802.11n Compliant MIMO Baseband and MAC Processor,*" ISSCC 2007*

High spectral efficiency comes at the cost of poor energy efficiency

Po y

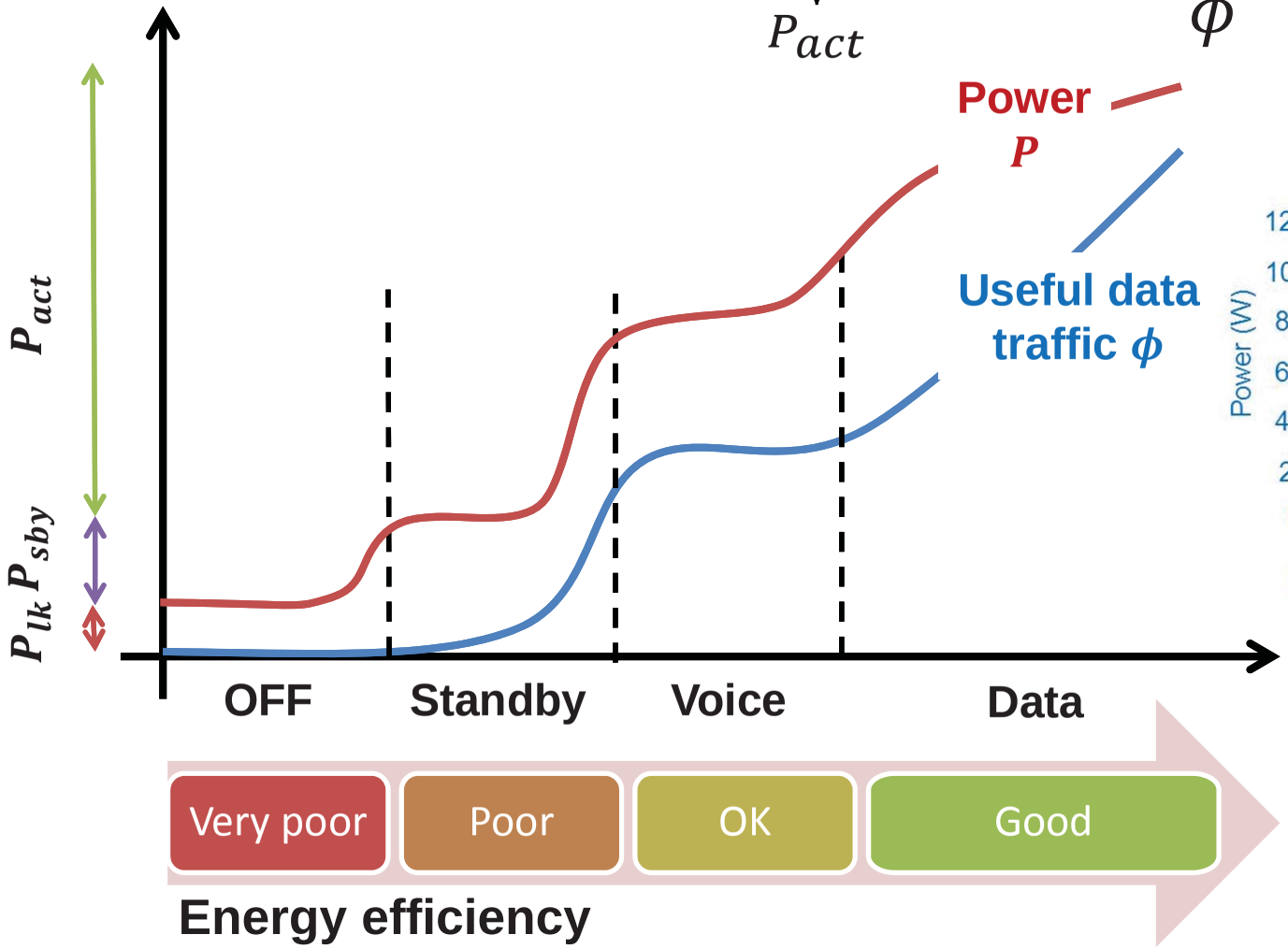
- Deter e

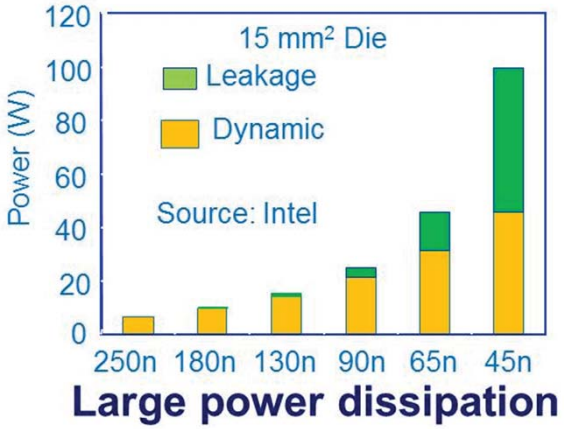
and standby activity

power consumption

Doing Nothing Well is Sometimes Difficult…

$$P = P_{lk} + P_{sby} + \underbrace{E_{act}\phi}_{P_{act}} \qquad E = \frac{P_{lk}}{\phi} + \frac{P_{sby}}{\phi} + E_{act}$$

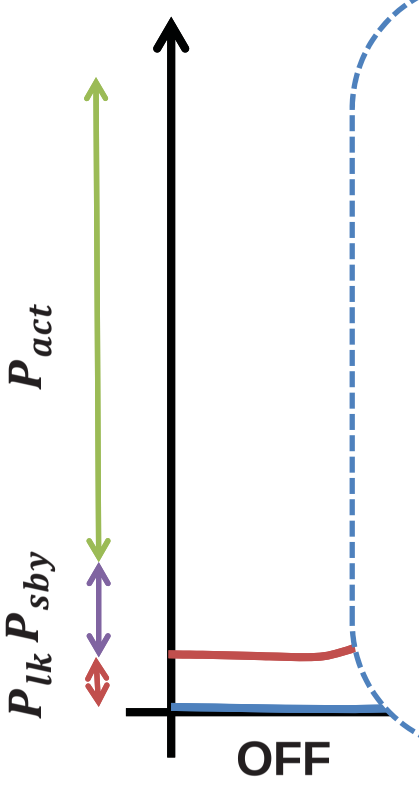




Large power dissipation

Doing Nothing Well is Sometimes Difficult…

$$P = P_{lk} + P_{sby} + E_{act}\phi \qquad E = \frac{P_{lk}}{} + \frac{P_{sby}}{} + E_{act}$$



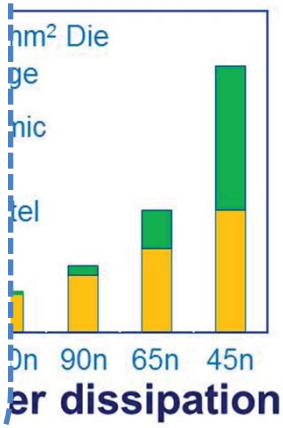
Leakage and standby currents dominate as

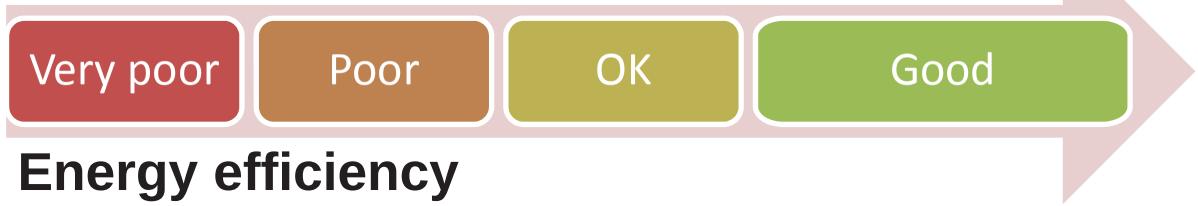
- **DSP becomes more energy efficient**
- **Workload decreases (e.g., standby)**

Low energy when "doing little" is "good enough"

Use high energy when need to "work hard"

Challenge: Energy Proportionality

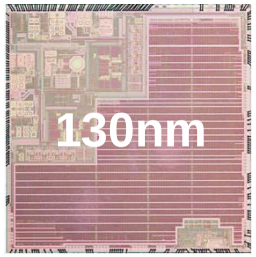




Should we Care about the Baseband Processing?

Lets check two examples

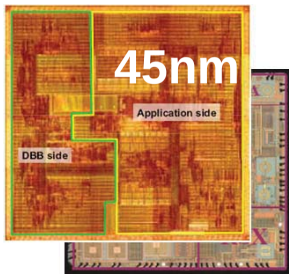
- 2x2 MIMO WLAN (IEEE 802.11n)



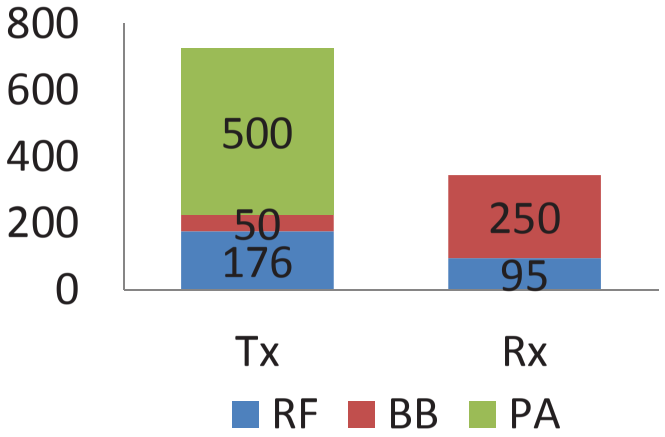
  - >70% area covered by baseband signal processing
  - DSP consumes significant power compared to RF (especially RX)

S. G. Sankaran, *et al.*, "Design and Implementation of a CMOS 802.11n SoC," *Comm. Magazine 2009*

- 3GPP Handset ASIC/MPSoC



  - >40% of the digital die covered by baseband signal processing
  - RX: Baseband consumes most of the total power



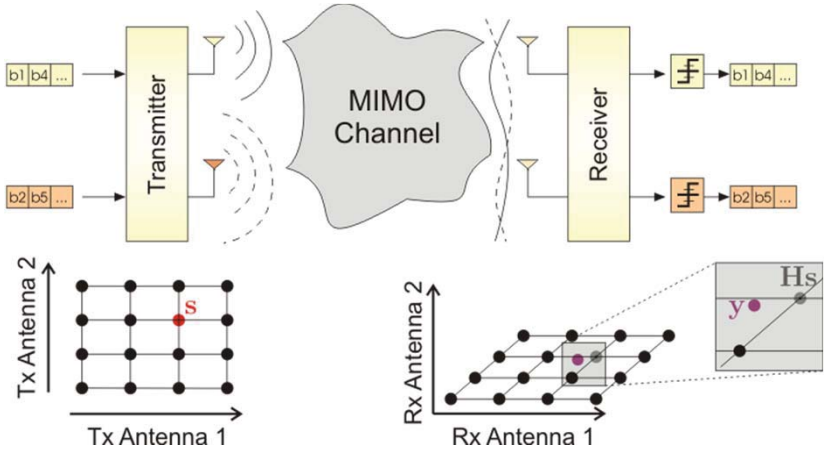
S. Kunie, et al., "Low power architecture and design techniques for mobile handset LSI Medity M2," ASP-DAC, 2008

Is there still a need for research?

MIMO Communication

MIMO: Transmit multiple data streams concurrently in same frequency band

- Used in almost all important standards



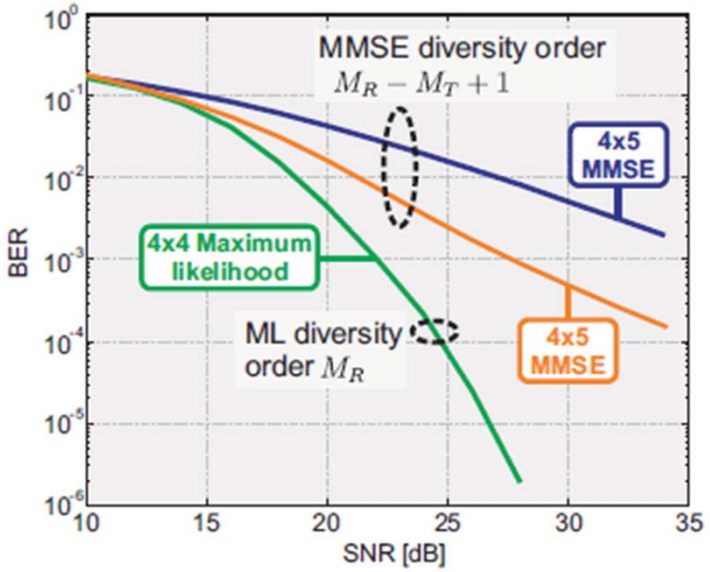
Task of the MIMO detector:

- Separation of multiplexed data streams
- Choice of the MIMO detector has significant impact on performance

Optimum MIMO detection:

$$\hat{\mathbf{s}} = \underset{\mathbf{s}\in\mathcal{O}^{M_T}}{\arg\min} \|\mathbf{y} - \mathbf{Hs}\|^2$$

- Straightforward solution: Check all candidates
- Number of candidates: exponential in spectral efficiency

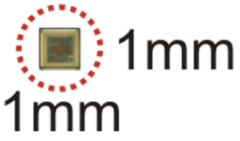


MIMO detection – a challenge of exponential complexity

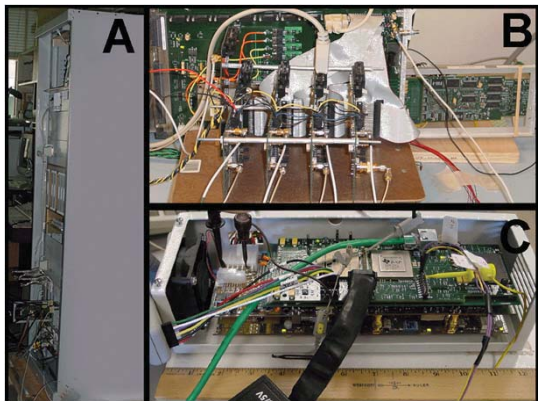
2002: 4-stream MIMO over UMTS

- Spectral efficiency 8 bits/s/Hz
- Examine **256 candidates**
  4 million times per second

1mm²
of silicon in a
250nm process



Source: Bell Labs Wireless Research, Holmdel, NJ



2002 : MIMO over UMTS with
1 Mbps for 31 users (8 bits/s/Hz)

2009: MIMO-WLAN

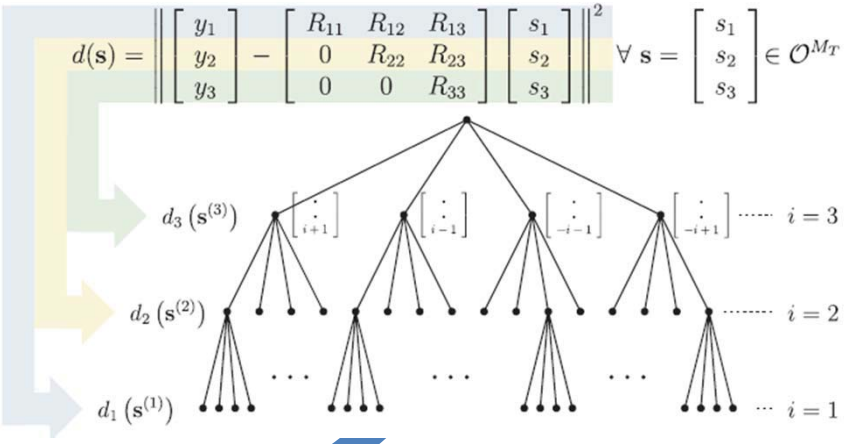
- Spectral efficiency 24 bits/s/Hz
- Examine **2^{24} candidates**
  40 million times per second

2009 : MIMO WLAN 600 Mbps
(24 bits/s/Hz)

New Algorithms Require New Architectures

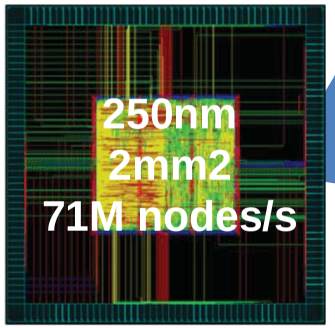
Sphere decoding

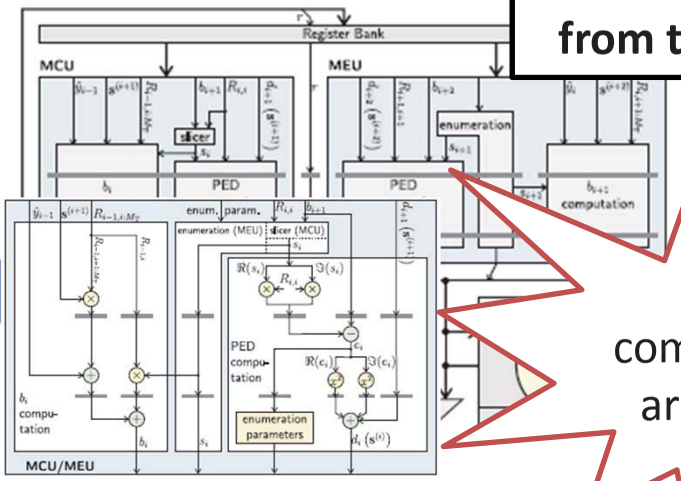
- Map the problem to a tree search
- Use branch and bound strategy for complexity reduction
- STS-sphere decoding provides soft-information for channel decoder



Tree-search is very different from typical DSP algorithms

2007 : STS Soft-output sphere-decoding with 10-40 Mbit/s



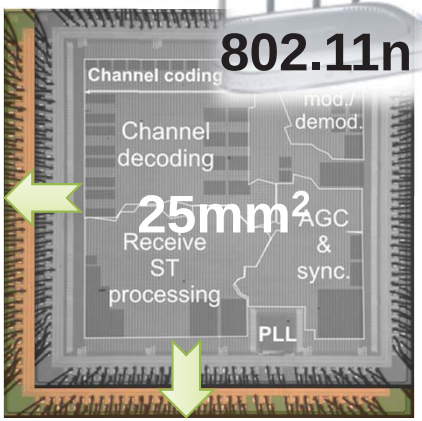


Requires completely new architectures

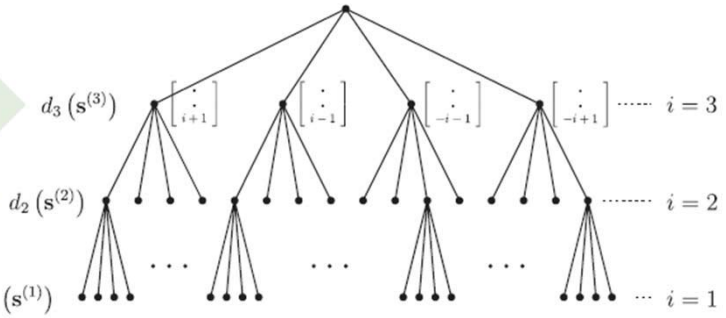
New Algorithms Require New Architectures

4 parallel instances
work at at 320MHz
➔ 1.28G nodes/s

2mm
2mm



$$d(\mathbf{s}) = \left\| \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} - \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ 0 & R_{22} & R_{23} \\ 0 & 0 & R_{33} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix} \right\|^2 \quad \forall\, \mathbf{s} = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix} \in \mathcal{O}^{M_T}$$

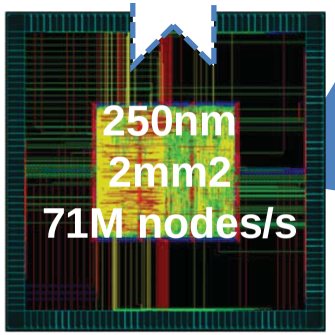


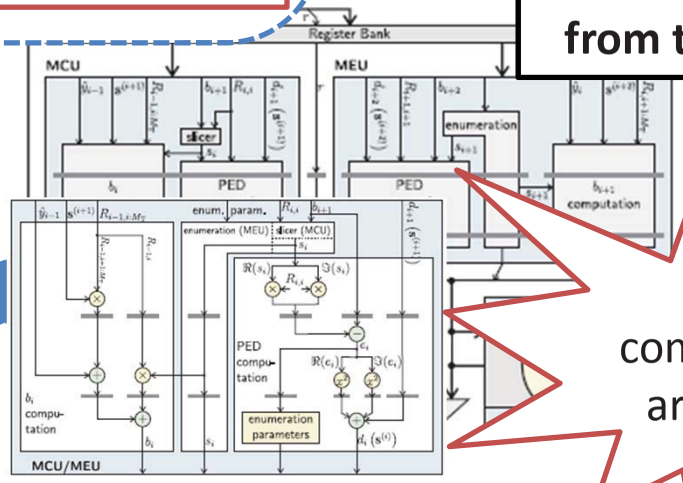
Near-optimum performance @ 600Mbps

Tree-search is very different from typical DSP algorithms

2007 ... Soft-output

Technology shrink & architecture optimization

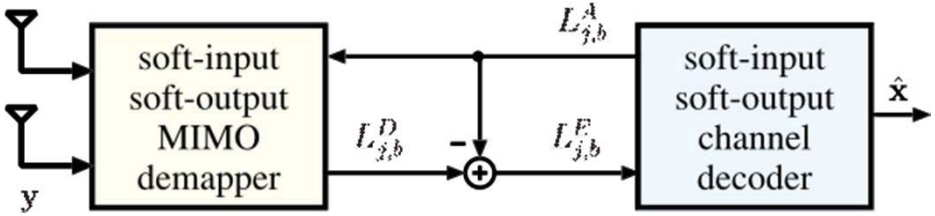


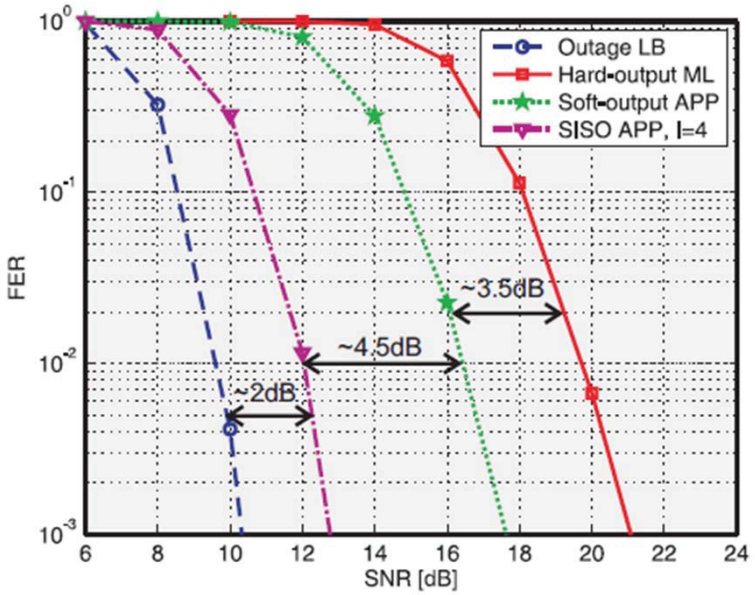


Requires completely new architectures

The Holy Grail: Iterative Detection and Decoding

- Exchange reliability information between MIMO detector and channel decoder
- Converge to optimum solution in multiple iterations



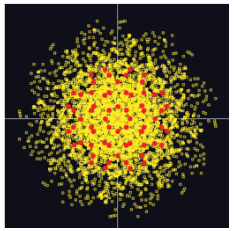


- Iterations require soft-in soft-out MIMO detection, which is even more complex
- Complexity for *N* iterations increases at least *N*-fold

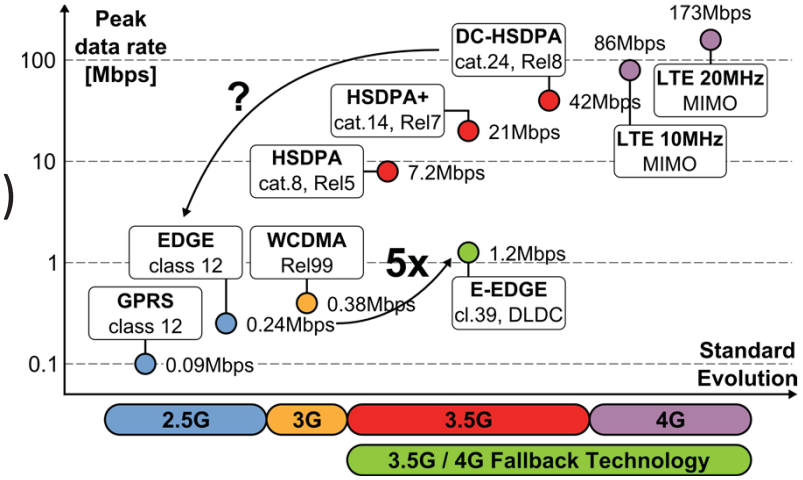
Spectral Efficiency Renders Equalization Difficult

3GPP 2007: Extension of 2G system GSM / EDGE toward higher data rates

- Higher modulation order (16QAM and 32QAM)
- 1.2x higher symbol rate
- Bandwidth remains unaltered



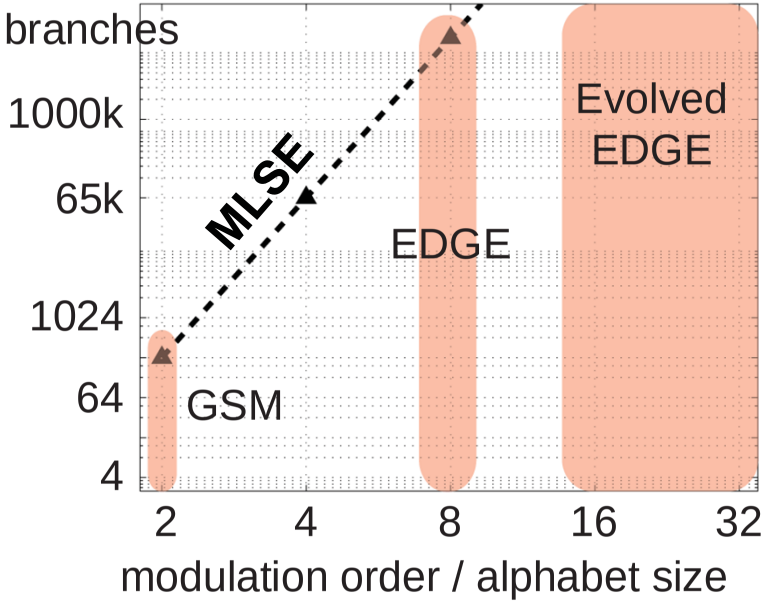
32QAM Tx-signal (Evolved EDGE)



Strong need for equalization

Optimum receiver: Maximum likelihood sequence estimation (MLSE)

- Complexity grows exponentially in spectral efficiency and channel length

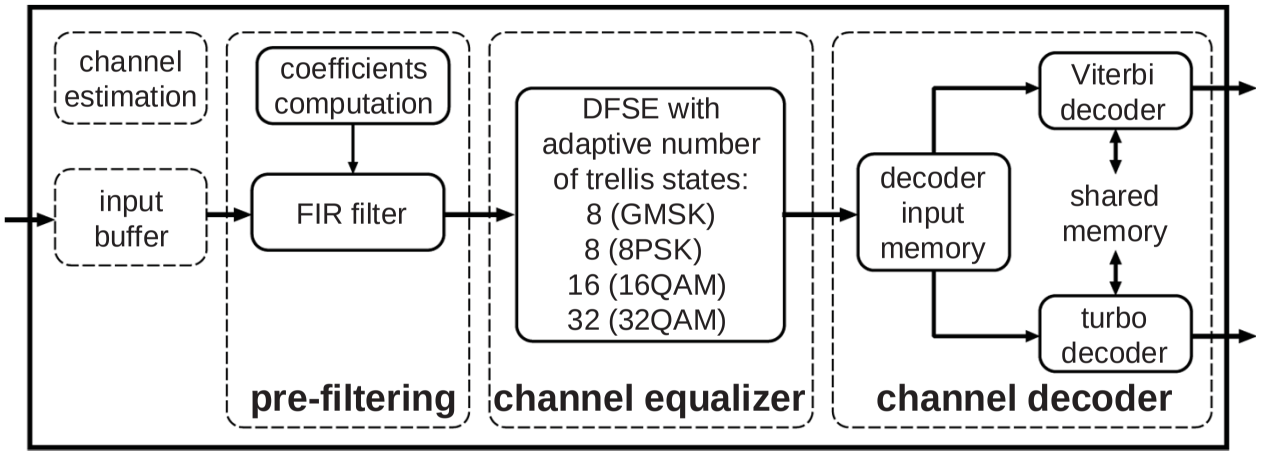


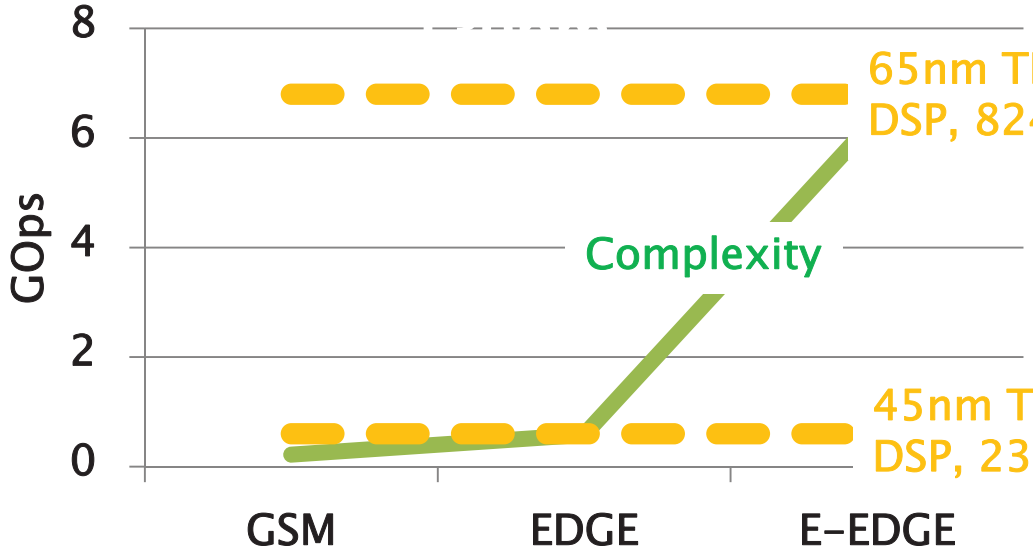
Impractical even in a 32nm process

Even Optimized Algorithms too Complex for SDR

Solution: channel shortening with decision feedback sequence estimation

- Orders of magnitude complexity reduction



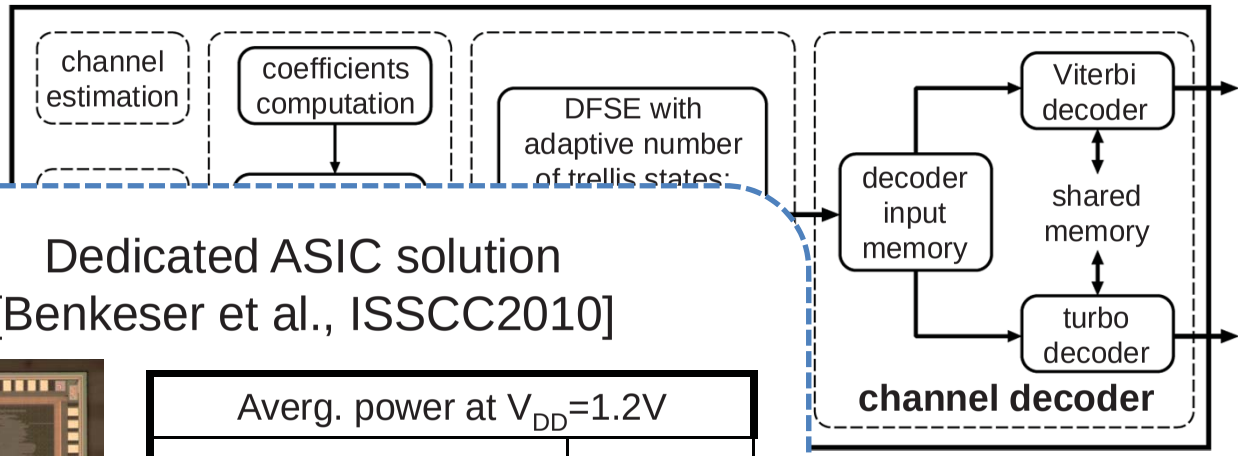


Complexity exceeds DSP capabilities even with advanced process nodes

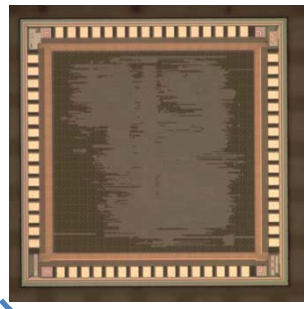
Even Optimized Algorithms too Complex for SDR

Solution: channel shortening with decision feedback sequence estimation

- Orders of magnitude complexity reduction



Dedicated ASIC solution
[Benkeser et al., ISSCC2010]



Averg. power at V_{DD}=1.2V	
EDGE (8PSK, CC)	**6.8mW**
E-EDGE (16QAM, TC)	**11.2mW**
E-EDGE (32QAM, TC)	**19.9mW**

exity exceeds DSP capabilities even with advanced process nodes

Prime Example for Complexity: Channel Decoding

Example: Low Density Parity Check Decoder

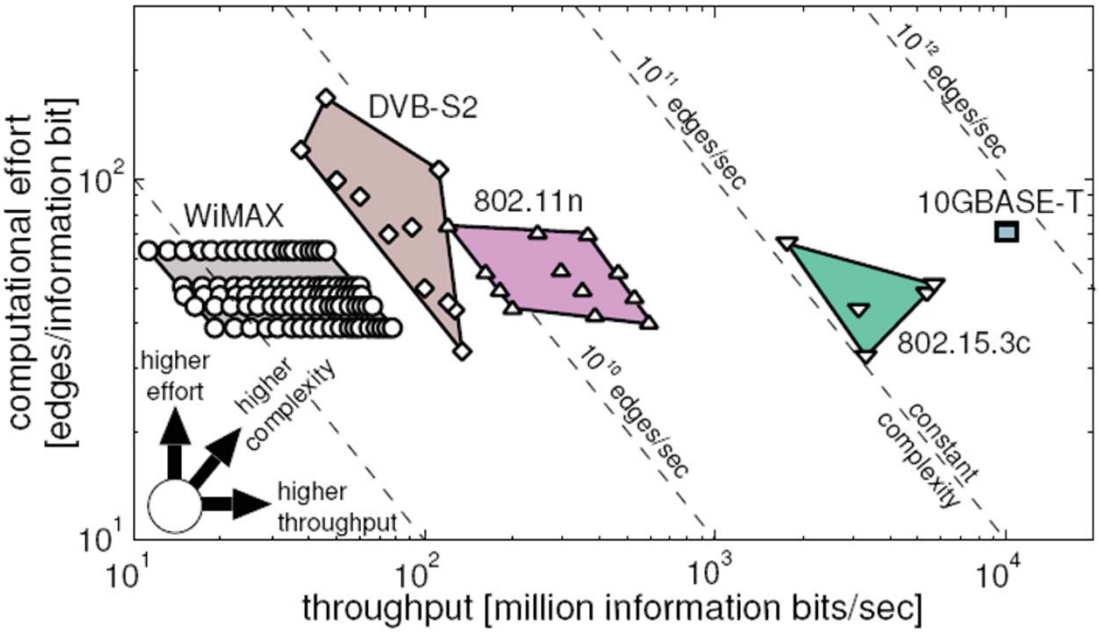
- **1962**: invented by R. G. Gallager
  - Performance close to the Shannon limit (on-par with Turbo codes)
  - Initially considered to complex for economic implementation
- **1999**: re-discovered by MacKay and Neal
  - VLSI technology allowed for the implementation of LDPC codes
- **Today**: LDPC codes are optional or mandatory in almost all relevant standards

LDPC Decoding is in fact Ideal for VLSI

Iterative message passing

- Large number of identical computational units, operating in parallel ➔ exploits resources available from scaling

- Different standards use different codes
- Different codes required within each standard
- Computational effort across standards spans 3 orders of magnitude

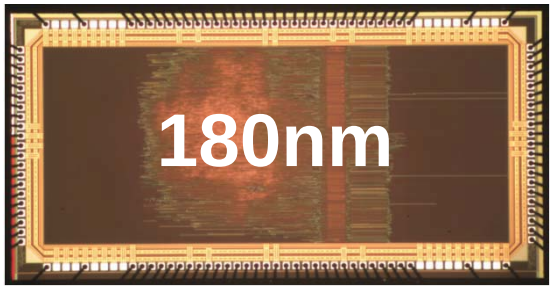


- Computational effort per bit remains almost constant

Technology Scaling Reduces Power

Reference design: LDPC decoder for IEEE 802.11n

- 208 MHz clock frequency
- 780 Mbps throughput
- 3.4mm^2 silicon area
- Workload ~50-100 GOps



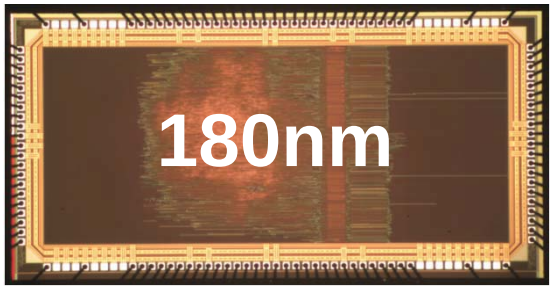
3.9 nJ/bit

2.3W @ 600Mbps

Technology Scaling Reduces Power

Reference design: LDPC decoder for IEEE 802.11n

- 208 MHz clock frequency
- 780 Mbps throughput
- 3.4mm^2 silicon area
- Workload ~50-100 GOps

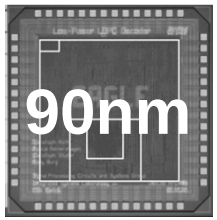


3.9 nJ/bit

2.3W @ 600Mbps

Constant throughput

6.4x better energy efficiency



- **Max. throughput** almost **doubles** with **half** silicon **area**

600 pJ/bit

360mW @ 600Mbps

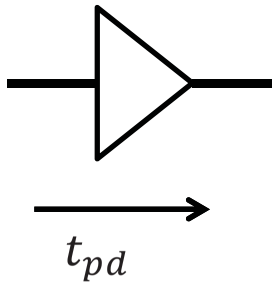
Technology scaling provides significant energy savings

Can we do still better??

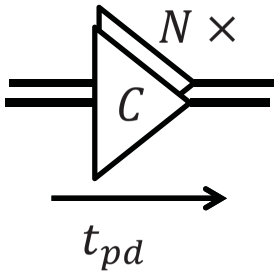
A Receipe for Energy Efficiency

Voltage Frequency Scaling: make things worse to make them better

- Design a circuit that works faster than planned (e.g., by replication)
  - When running at the same speed and voltage, energy efficiency becomes worse

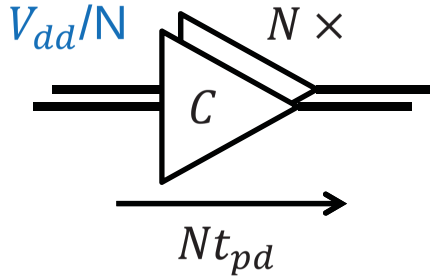


$$E = \frac{\alpha}{2} C V_{dd}^2$$



$$E = \frac{\alpha}{2} NC V_{dd}^2$$

- Utilize the fact that $t_{pd} \propto V_{dd}$
  - Reduce voltage until it just meets the delay constraint

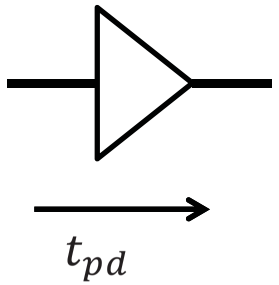


$$E = \frac{\alpha}{2} NC \frac{V_{dd}^2}{N^2} = \frac{\alpha}{2N} V_{dd}^2$$

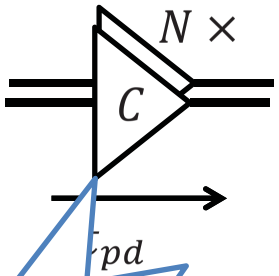
A Receipe for Energy Efficiency

Voltage Frequency Scaling: make things worse to make them better

- Design a circuit that works faster than planned (e.g., by replication)
  - When running at the same speed and voltage, energy efficiency becomes worse



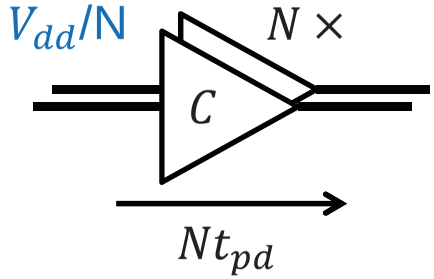
$$E = \frac{\alpha}{2} C V_{dd}^2$$



$$E = \frac{\alpha}{2} NC V_{dd}^2$$

- Utilize the fact that $t_{pd} \propto V_{dd}$
  - Reduce voltage until it just meets the c… st… nt

1/N better energy efficiency

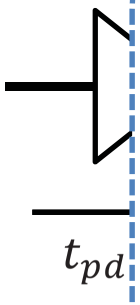


$$E = \frac{\alpha}{2} NC \frac{V_{dd}^2}{N^2} = \frac{\alpha}{2N} V_{dd}^2$$

A Receipe for Energy Efficiency

Voltage Frequency Scaling: make things worse to make them better

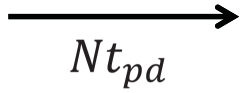
- Design a circuit that works faster than planned (e.g., by replication)
  - When ru becomes worse

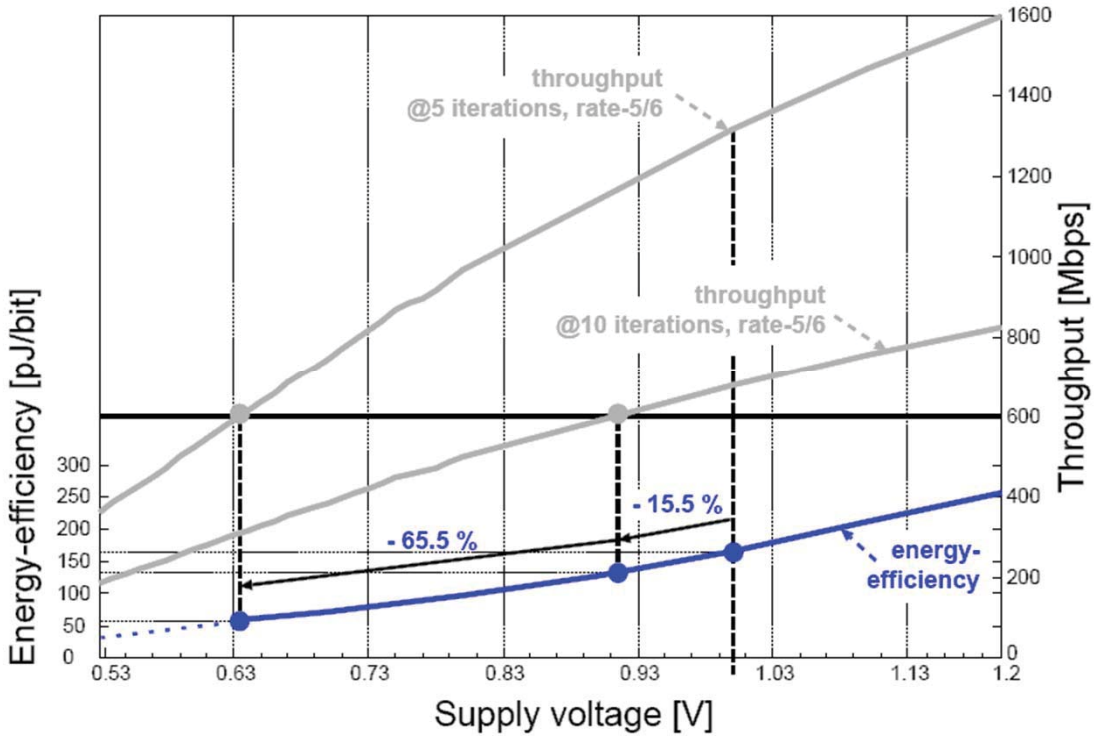


NCV_{dd}^2

- Utilize th
  - Reduce v

dd





The End of Moore's Law

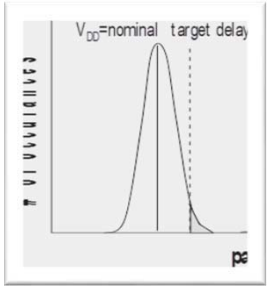
Moore's Law has enabled advances in Communications for 2-3 decades…..

What can communications and information theory do for technology?

Why are Chips Always Working?

Production test is needed

- Identify chips with production defects
- Classify functional dies according to the speed they can reach



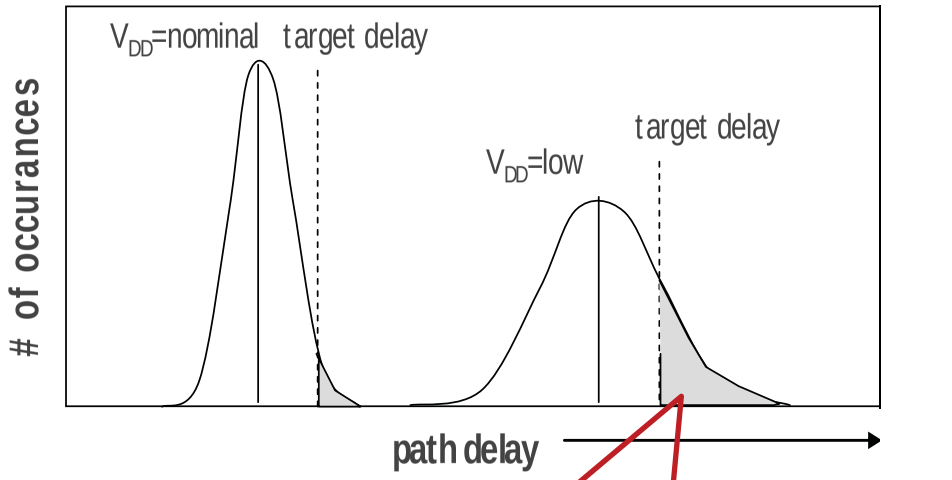
- Microprocessors: functional dies sold at different prices depending on their speed

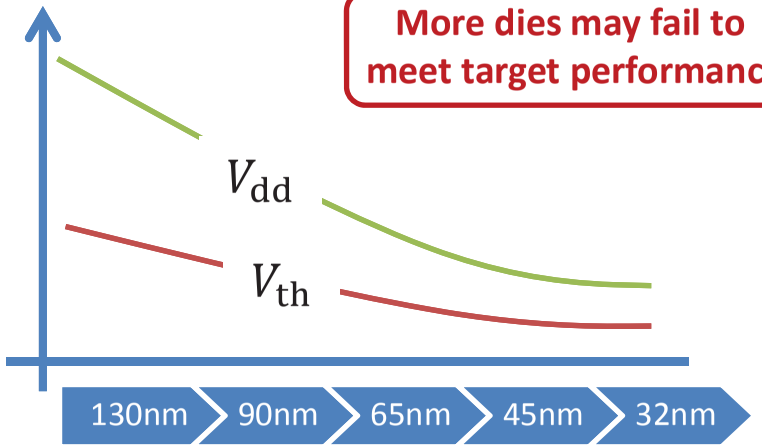
Speed binning improves yield

- Communication ASICs: need to run at a predefined fixed clock speed
  - Slow dies must be discarded
  - Fast dies do not exploit better performance

Variability Renders Voltage Scaling Difficult

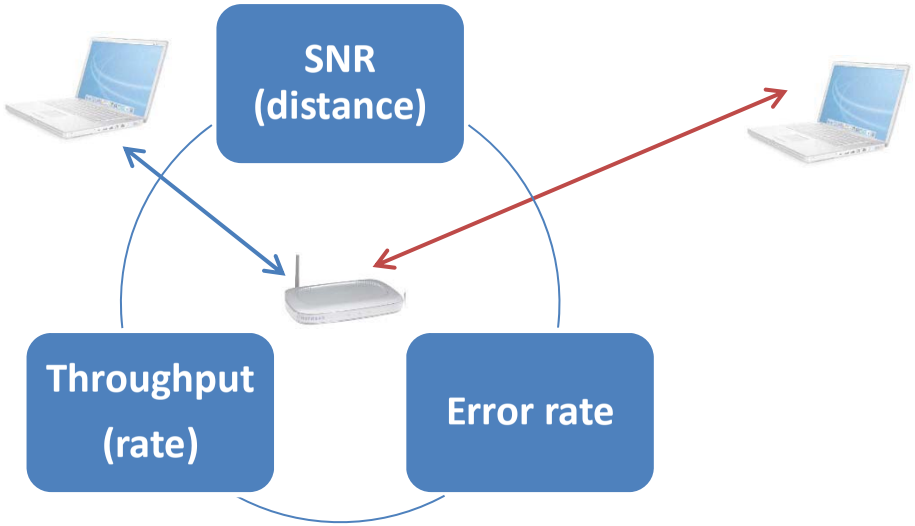
- Voltage Scaling
  - ✓ Quadratic power savings
  - x Increases mean delay making circuits slower
  - x Increases also delay variance making harder to meet target performance
- Conventional Solution
  - Overdesign: assume pessimistic guard bands (timing, voltage)
  - x Higher power consumption on average
  - x Limit the returns performance, power) from technology scaling

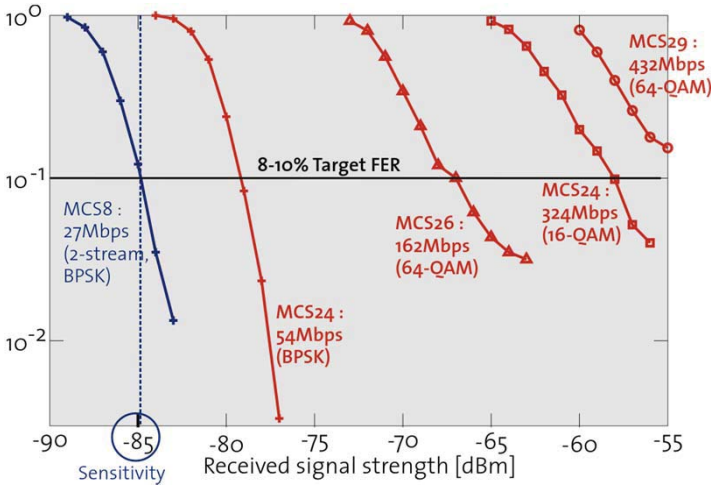


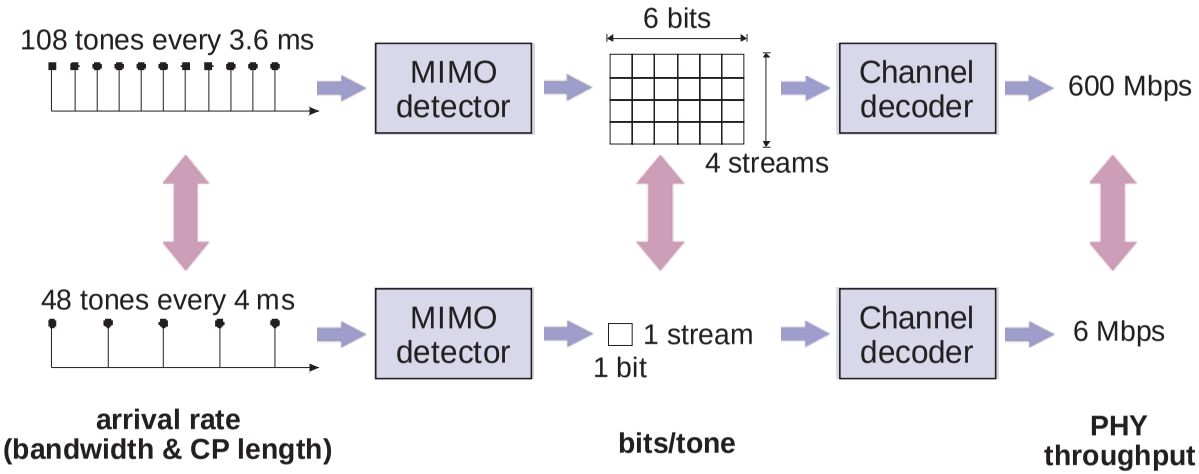


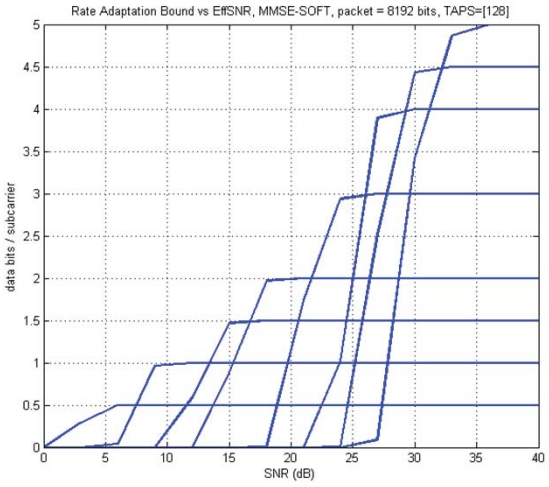
Supply voltage approaches the threshold voltage

Scalability at the System Level



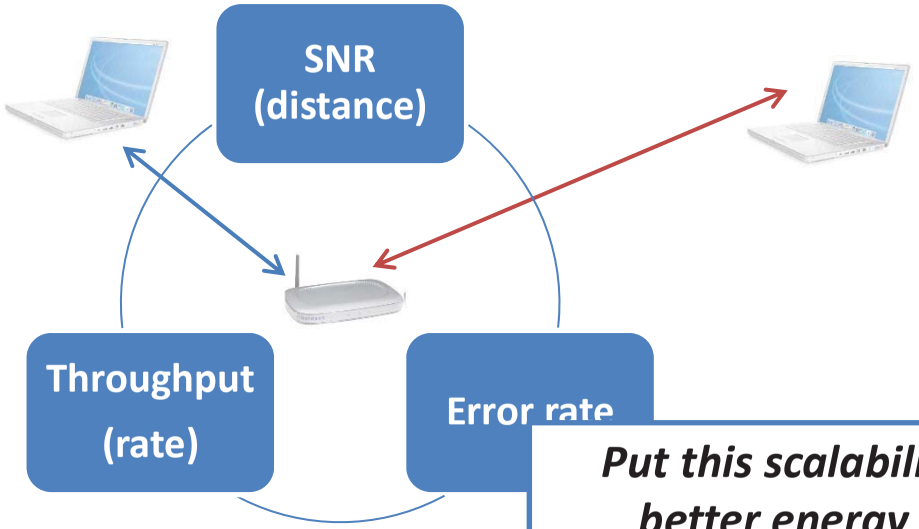


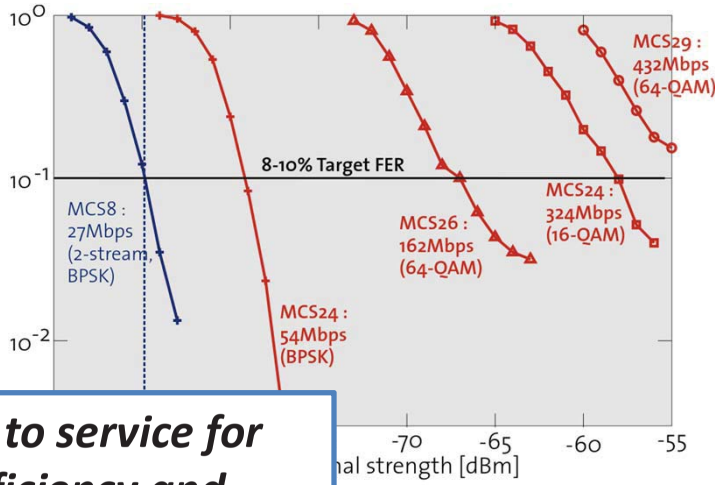




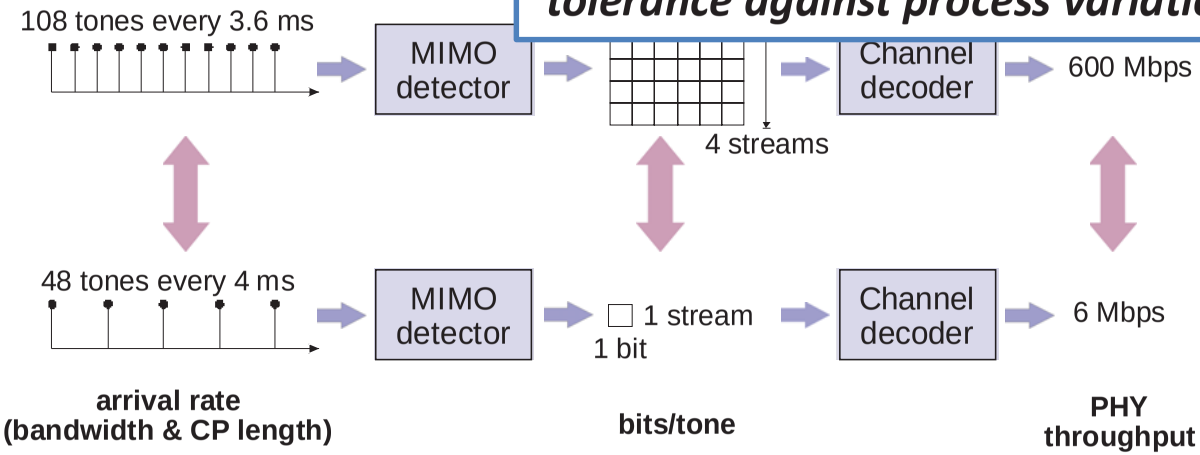
Rate adaptation is routinely used to deal with constantly varying channel conditions

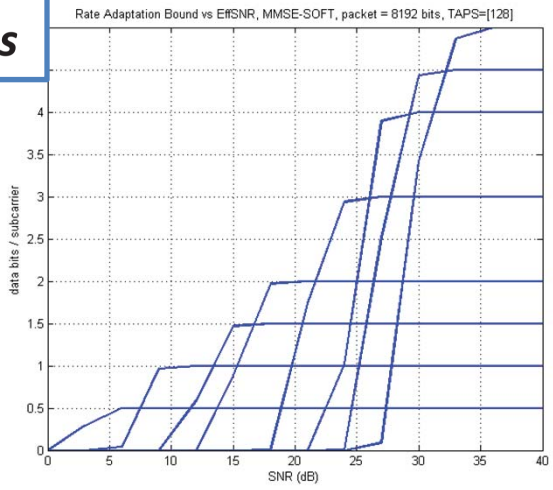
Scalability at the System Level





Put this scalability to service for better energy efficiency and tolerance against process variations



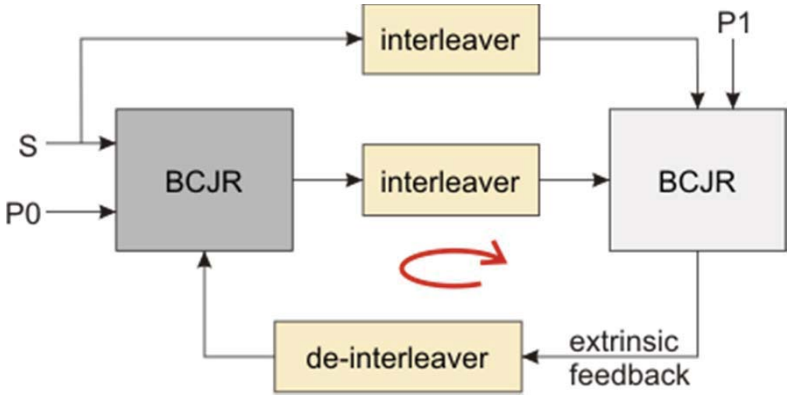


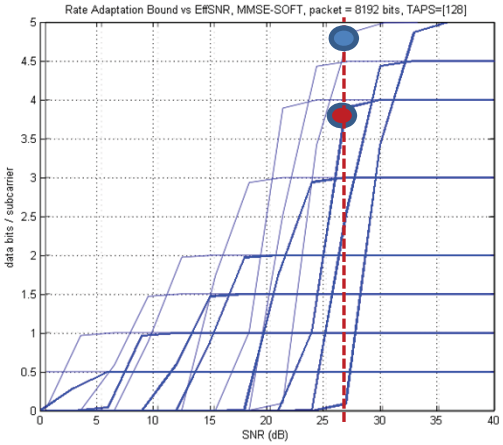
Rate adaptation is routinely used to deal with constantly varying channel conditions

Scalable Complexity Algorithms to the Rescue

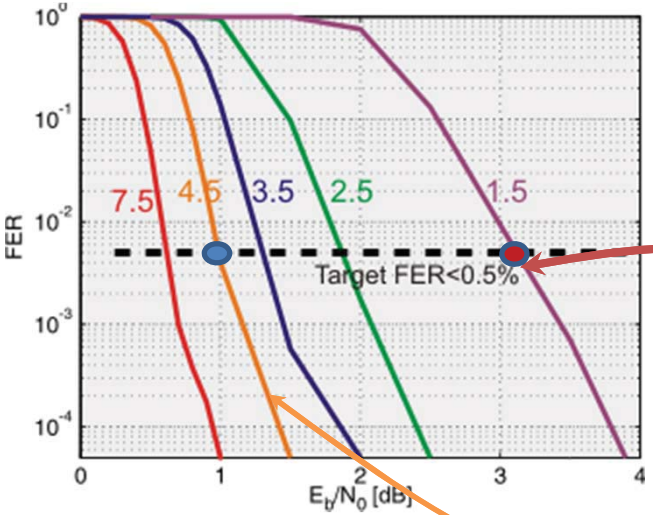
Iterative receivers / decoders: data passes multiple times through the same algorithm

- Performance improves with each iteratrion
- Deminishing returns after few iterations

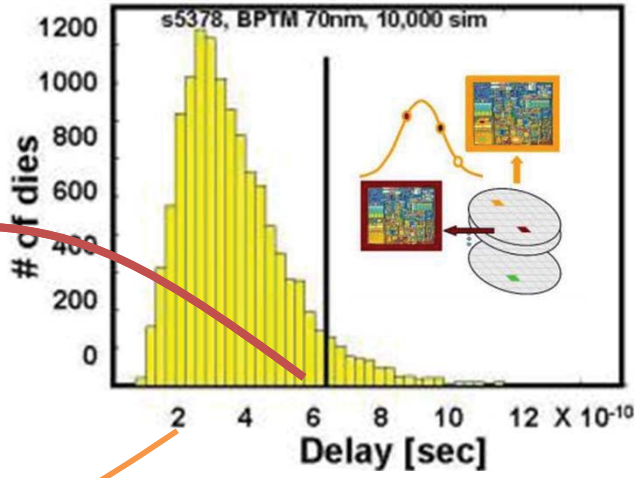




Achievable rate decreases



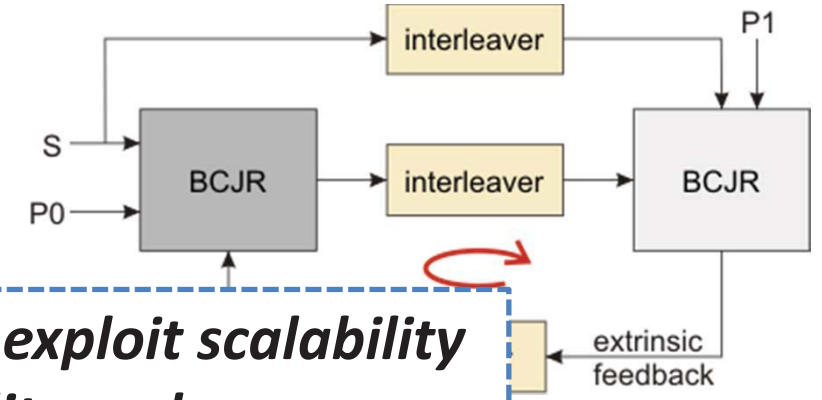
Achieve same rate only at a shorter distance



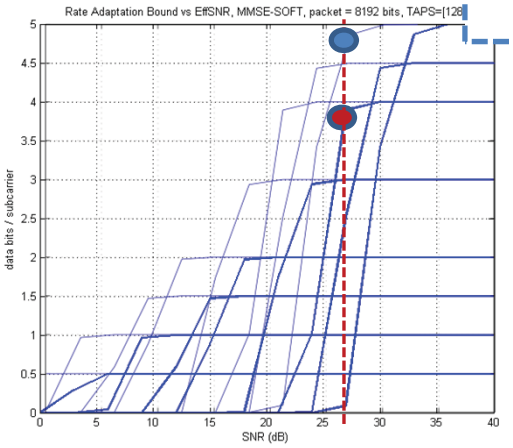
Scalable Complexity Algorithms to the Rescue

Iterative receivers / decoders: data passes multiple times through the same algorithm

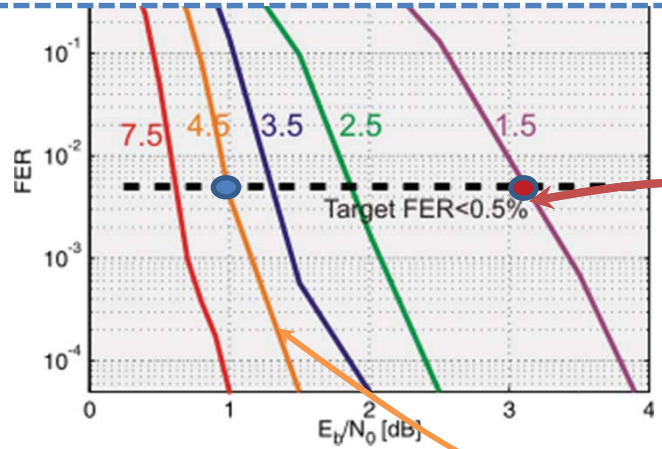
- Performance improves with each iteratrion
- Deminishing re



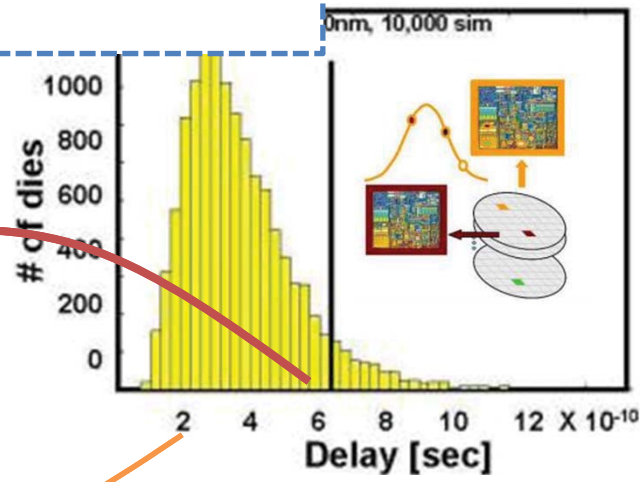
Yield improvement: exploit scalability to retain functionality under process variations



Achievable rate decreases

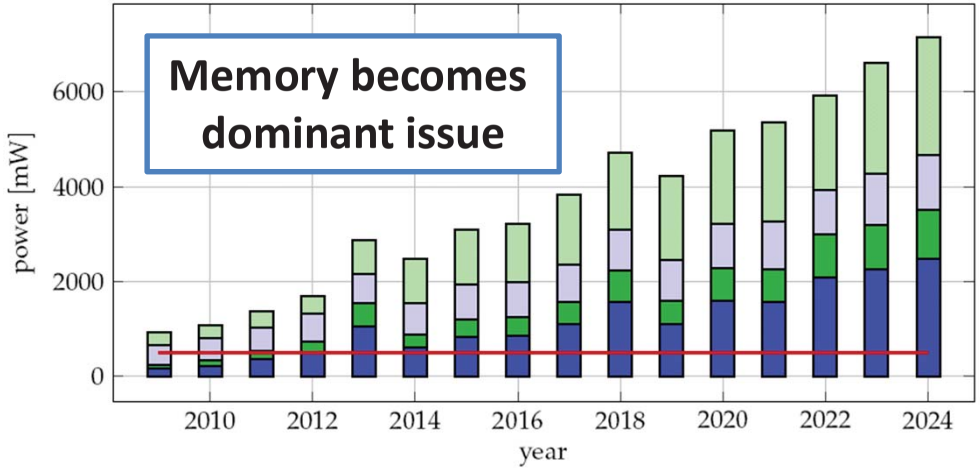


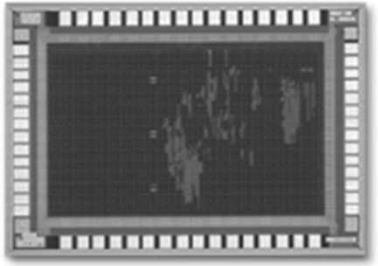
Achieve same rate only at a shorter distance



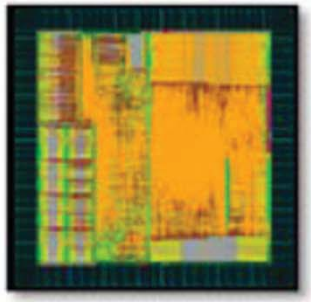
In Search for the Worst

- Baseband processor is comprised of logic and memory (on-chip and sometimes off-chip)
- Prediction from ITRS roadmap:

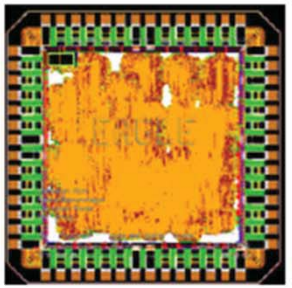




Turbo decoder
55% memory



Turbo equalizer
61% storage



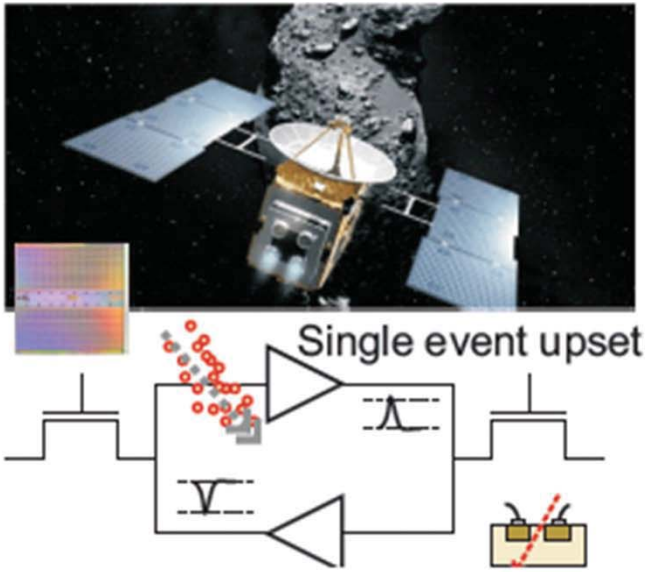
1.5 Gbps LDPC
>60% memory

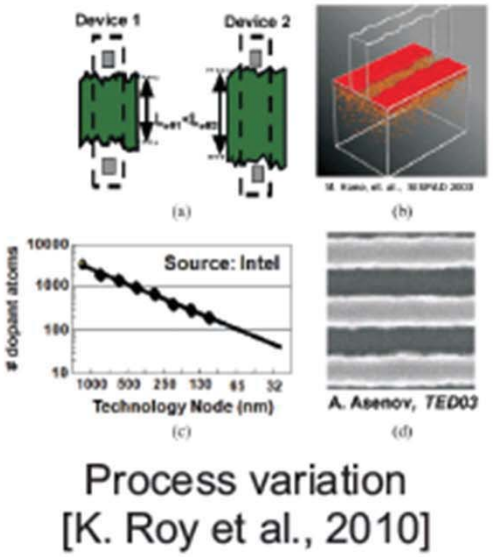
Primary concern: embedded (small- medium size) memories in DSP blocks

- Occupy a significant percentage of the area
- Consume a significant share of the power
- Memories are the primary source of failure (yield loss)

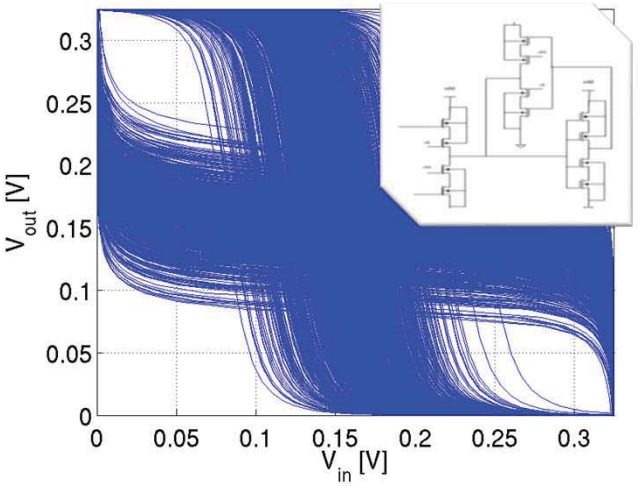
10 Gbps LDPC
25% memory

The End of Moore's Law





Process variation
[K. Roy et al., 2010]



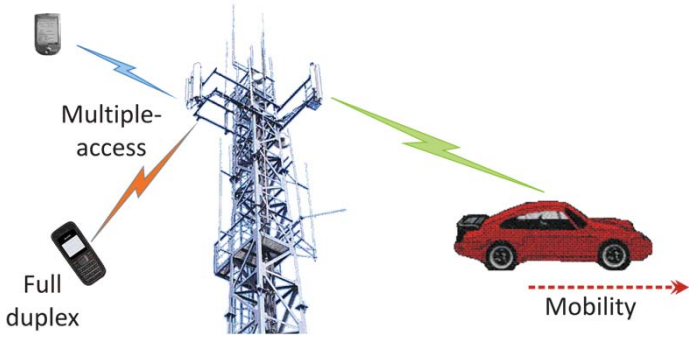
- Denser and larger memories are more susceptible to radiation
- Process variation leads to static errors and dysfunctional cells
- Reduced noise margins and supply noise induce errors in weak cells

Manufacturing circuits (memories) that are actually functional and robust becomes increasingly difficult

Inherent Error Resilience

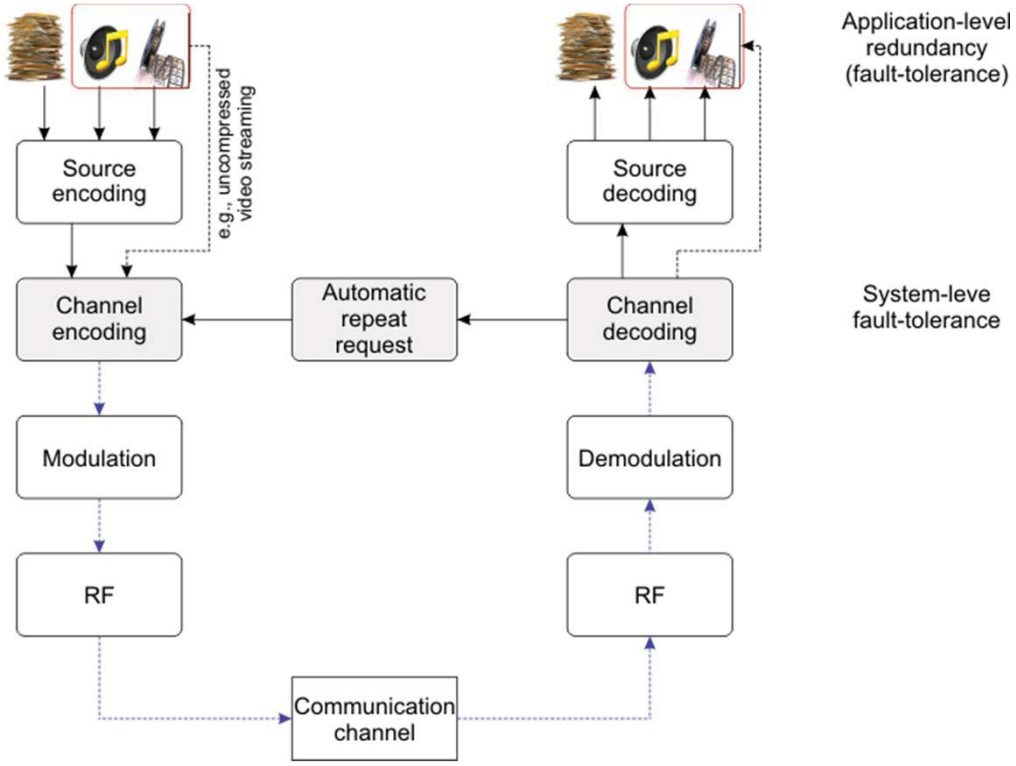
Fault-tolerant by design:

- Channel fading (random fluctuation of signal strength)
- Unknown (noisy) channel parameters
- Thermal noise and interference

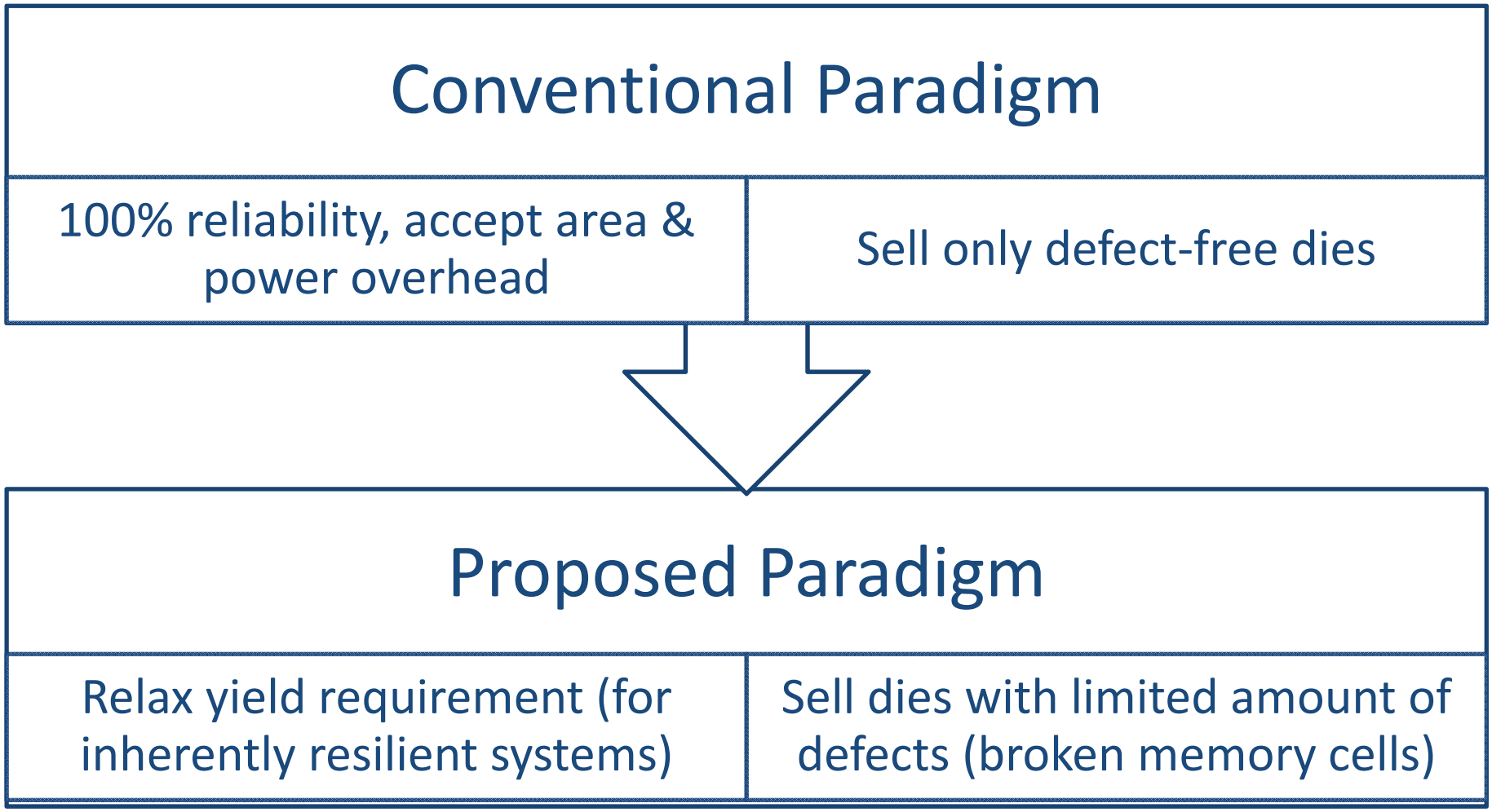


System-level mechanisms to restore reliable behavior:

- Forward error correction coding
- Automatic repeat request
- Application-level fault tolerance (e.g., video over UDP)



Design Paradigm Shift (or: Selling "Defect" Chips)



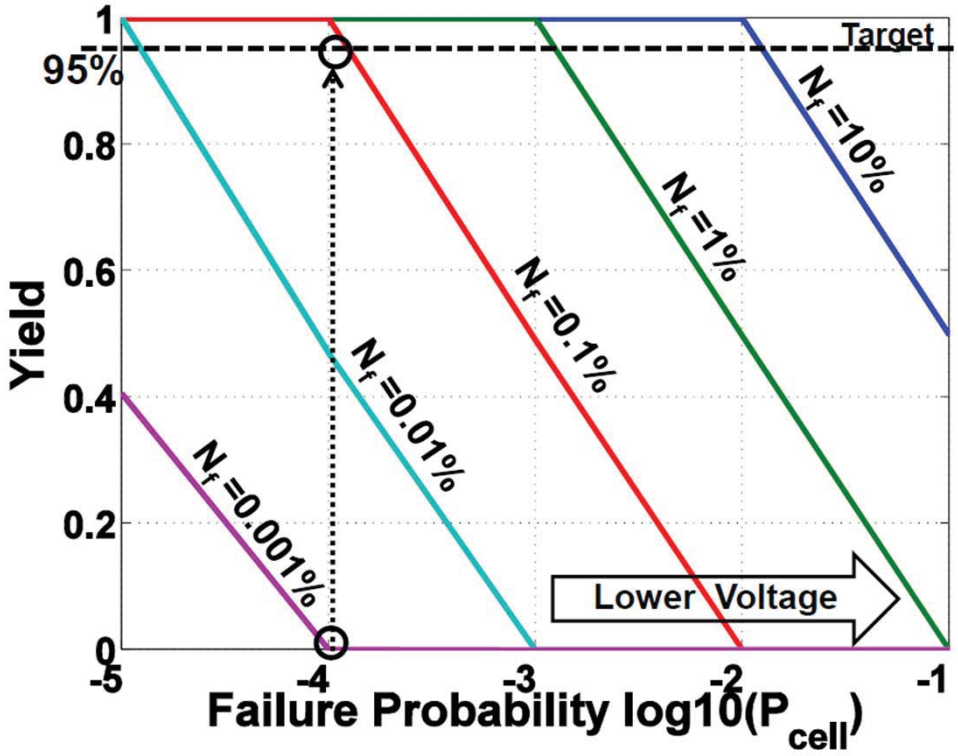
Relaxing the Yield Requirement

Conventional yield definition

- Accepting only chips with no defects

Proposed yield definition $Y(N_f)$ for systems with inherent hardware-error resilience

- Chips with at most N_f faulty memory cells pass inspection



Accepting more defects means

- **Higher yield,** and/or
- **Lower voltage & power**

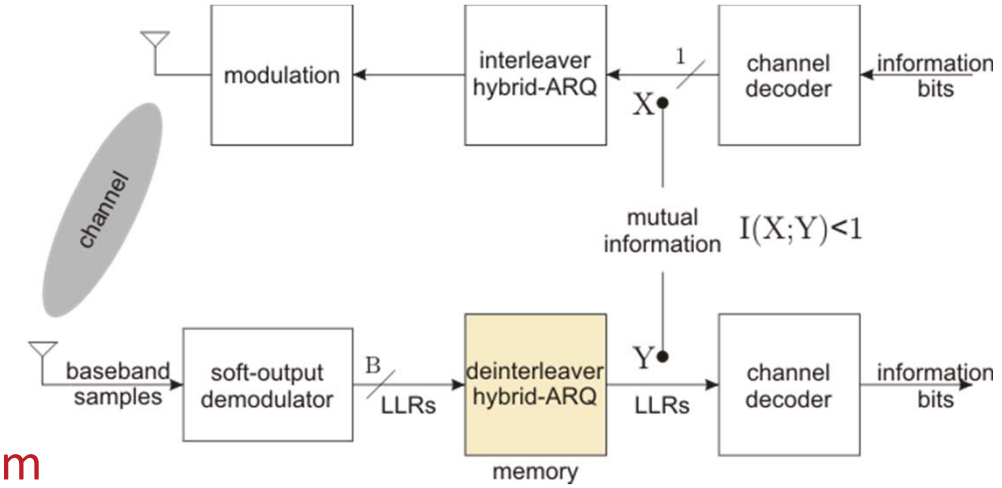
What is the impact on system-level metrics (throughput) ?

Impact of Unreliable Storage on Communication

Example : Communication system with bit-interleaved-coded modulation (BICM)

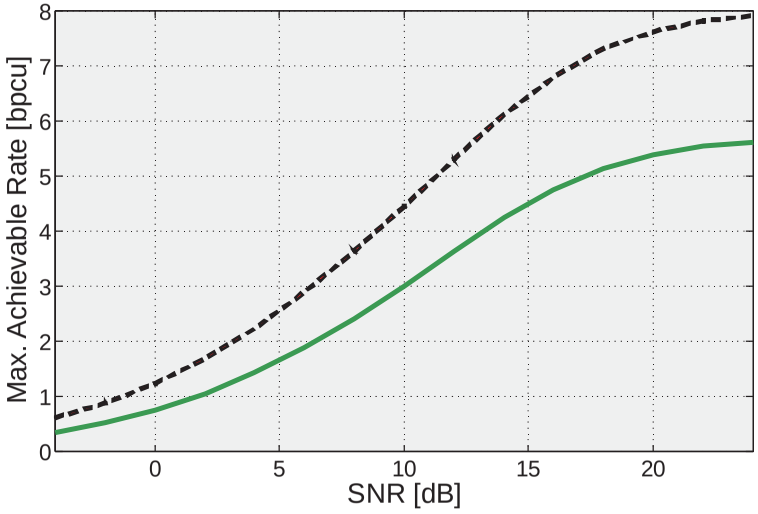
- HSPDA. WiMAX, 3GPP LTE, GSM, WLAN, …

Interleaver memory: stores reliability information of the received data bits



Fault model : de-interleaver built from unreliable memory (5% BER)

- Binary symmetric channel
- Randomized error locations

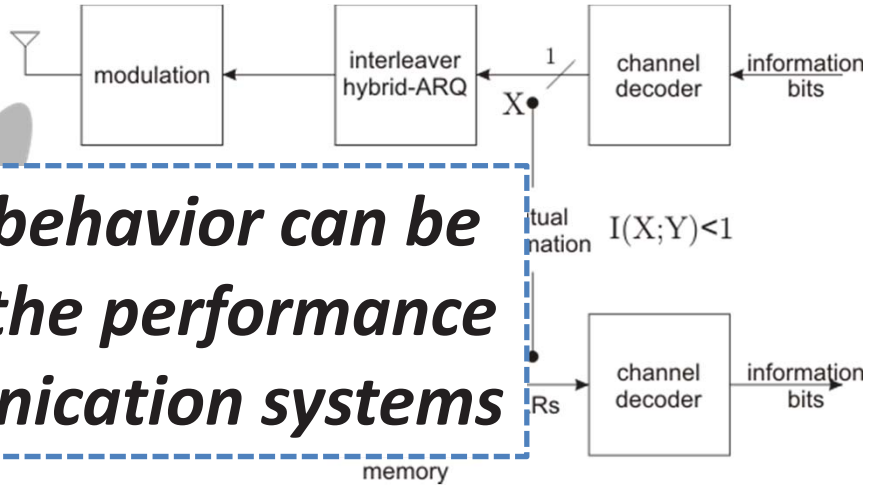


Collaboration with TU-Vienna (Matz, Novak)

Impact of Unreliable Storage on Communication

Example : Communication system with bit-interleaved-coded modulation (BICM)

- HSPDA. WiMAX, 3GPP LTE, GSM, WLAN, …

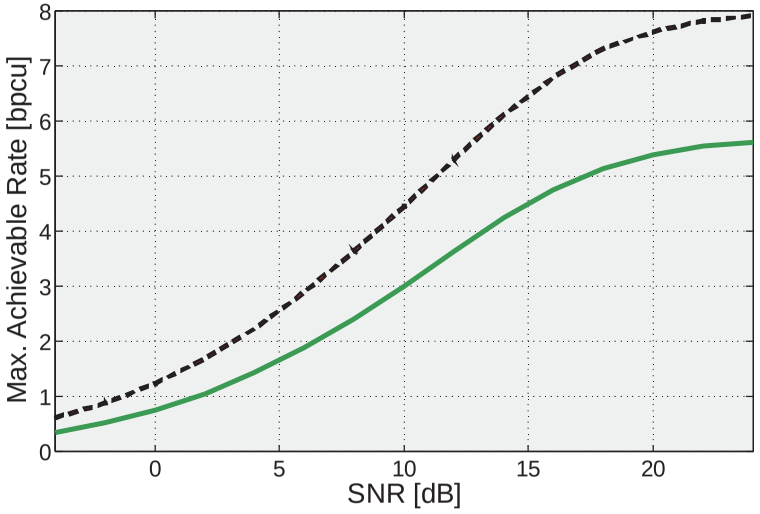


Interleaver memory: stores reliability i… receiv…

Unreliable circuit behavior can be incorporated into the performance analysis of communication systems

Fault model : de-interleaver built from unreliable memory (5% BER)

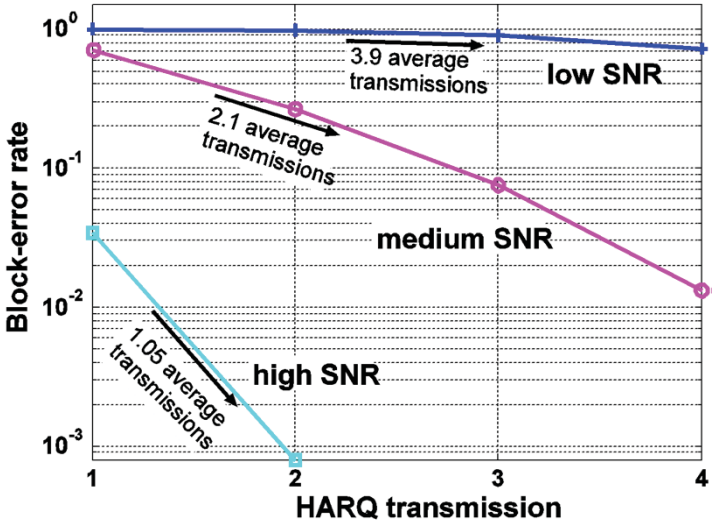
- Binary symmetric channel
- Randomized error locations

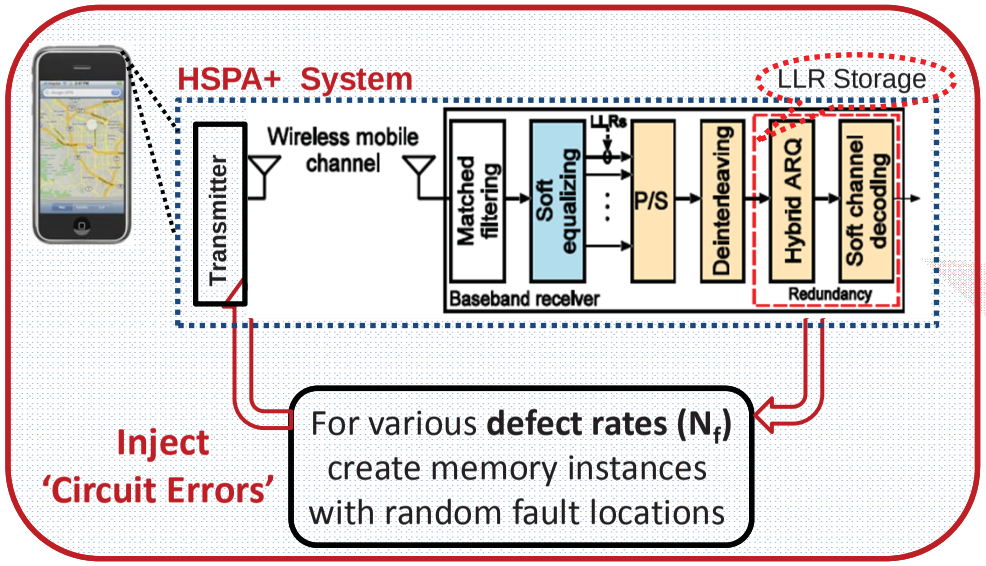


Collaboration with TU-Vienna (Matz, Novak)

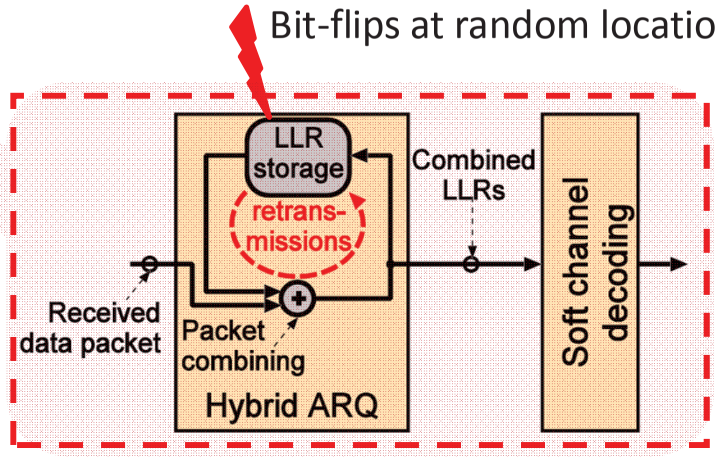
Resilience to Hardware Defects in HSPA+

- Explore the **resilience limits** of wireless communication systems to hardware defects
- **Simulation of complete HSPA+ system**, with error injection (in HARQ memory)

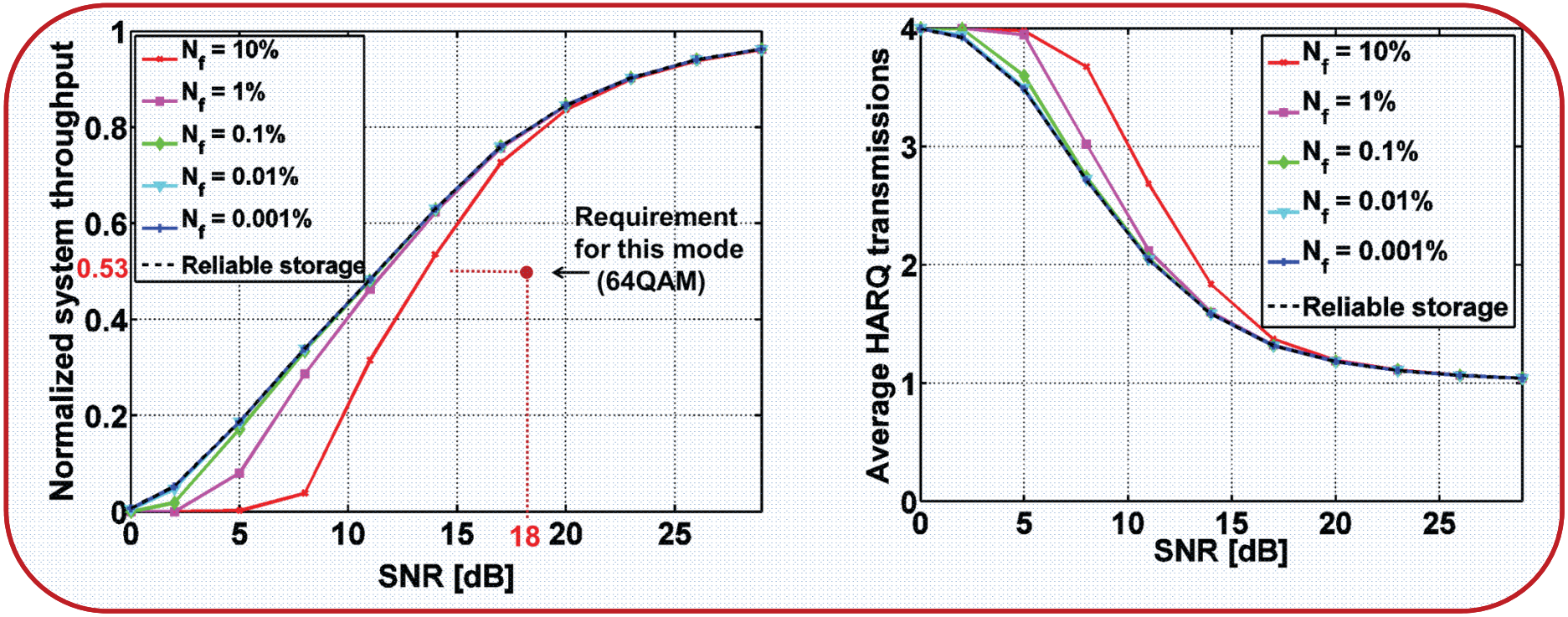




HARQ memory:
Bit-flips at random locations



Inherent Hardware-Error Resilience Limit



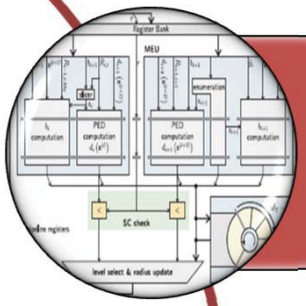
20 errors (N_f=0.01%, 200kb LLR storage)

- (Almost) same throughput as for defect-free hardware

2'000 errors (N_f=1%)

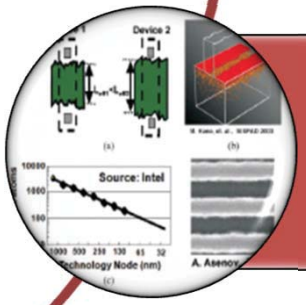
- Achieve required throughput (but clear penalty w.r.t. defect-free hardware)
- Power reduction by allowing low voltages (~200 mV less)

Despite Enough Transistors: Three Great Challenges



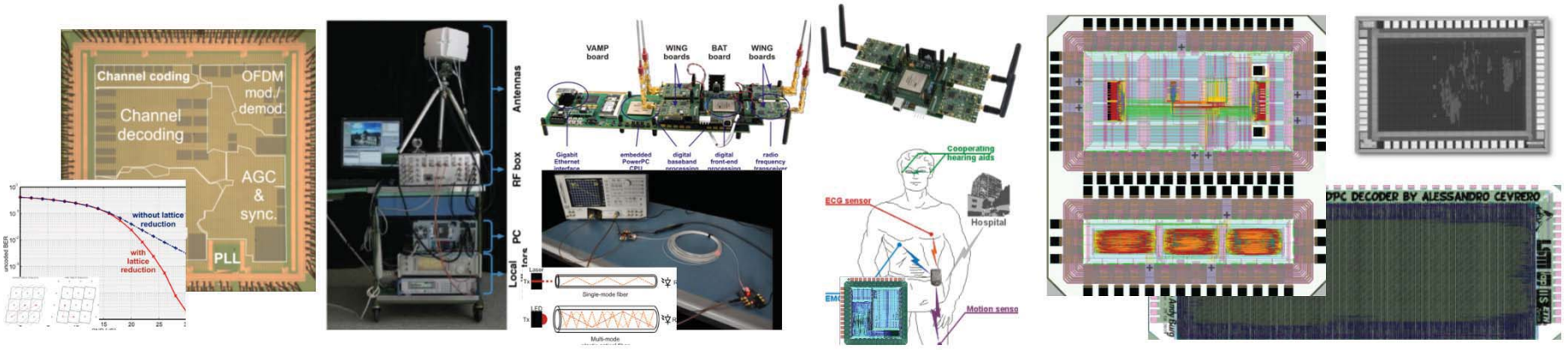
New Algorithms and Architectures for Bypassing the Exponential Complexity Associated with Spectral Efficiency

Improving Energy-Efficiency (nJ/bit) and Achieving Energy Proportionality in Communications



Exploiting System Level Error Tolerance to Cope with the Issues of Depp Submicron Integration

Thank you for your attention!



Signal processing algorithms: MIMO detection, sparse-channel estimation, equalization, CS-ADCs for spectrum sensing

System design and test (prototype implementations): MIMO, visible light communication, communication over plastic optical fibers, GSM/Evolved EDGE, TD-SCDMA

VLSI circuits for communications: circuit techniques for low-power and ultra-high speed signal processing, fault-tolerant signal processing for deep submicron VLSI, VLSI for embedded systems

andreas.burg@epfl.ch

http://tcl.epfl.ch/