"ATRS: an alternative roadmap for semiconductors, technology evolution and impacts on system architecture"

ASYNC'2006

March 14, 2006

Jean-Pierre.Schoellkopf@ST.com

Front-End Technology and Manufacturing (FTM)
Central CAD & Design Solutions(CCDS)
STMicroelectronics

OUTLINE

- One Challenge for the future:
  Process Variability
  (LASGA concept as a potential solution)

- Another Challenge for the future:
  Improve MOPS/Watt @ constant dynamic power
  (ATRS concept as a potential solution)

Part1: Challenge for the future:

Process Variability

Challenge for the future: Process Variability

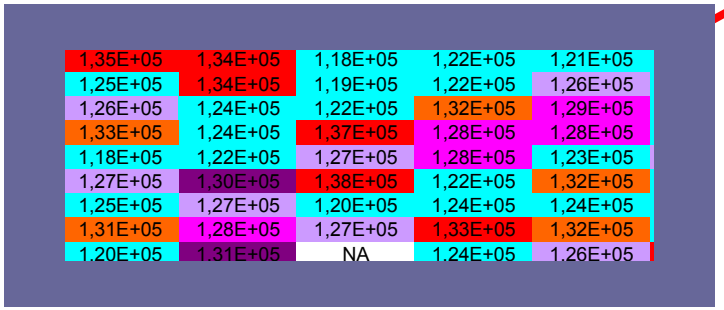
- **Intra-die variations increase:**
  - Implement LOCAL process compensations
  - Body Bias, Sourve Bias, Frequency tuning, VDD tuning
  - Statistical Design versus Worst Case
- **Measurement Techniques are not obvious:**
  - How to catch local variations ?
  - How insensitive are we to Temperature variations ?
- **Compensation Techniques are tricky:**
  - Require High precision Analog
  - Must be low power (must not consume more than what we could gain !)
- **Variation tolerant design techniques has to be deployed:**
  - Pure Asynchronous like QDI ?
  - Locally Adaptative Synchronous (see next slides)
  - Redundancy ?

Wafer Map 90nm:
Ring Oscillator Frequency

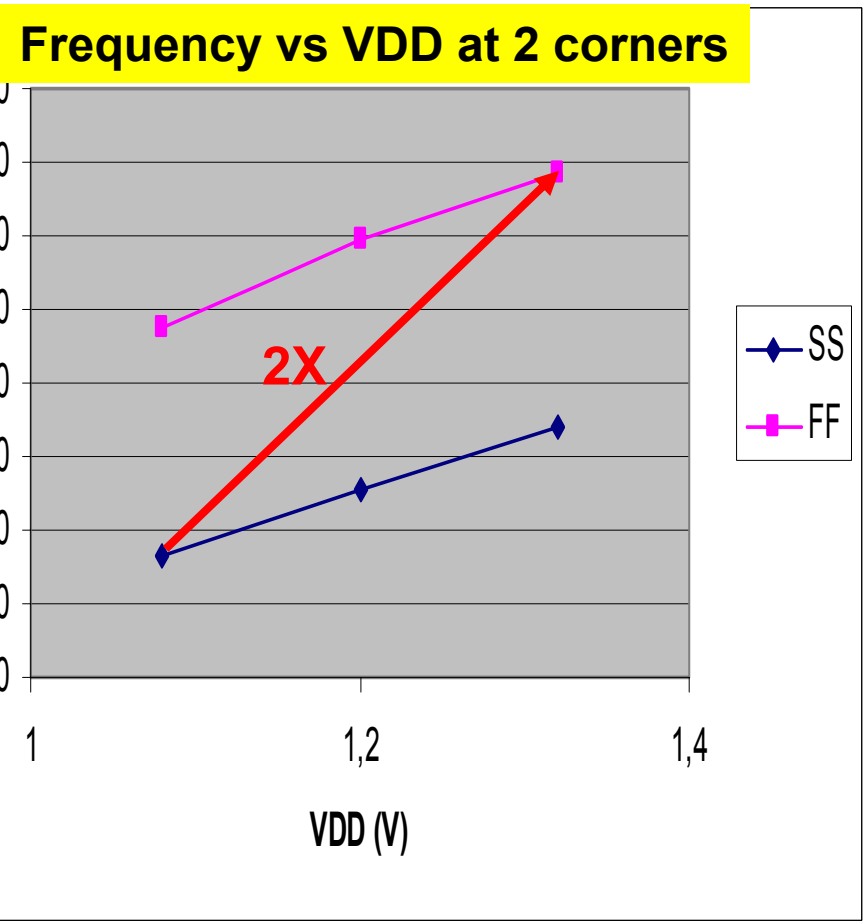
Process spread is increasing

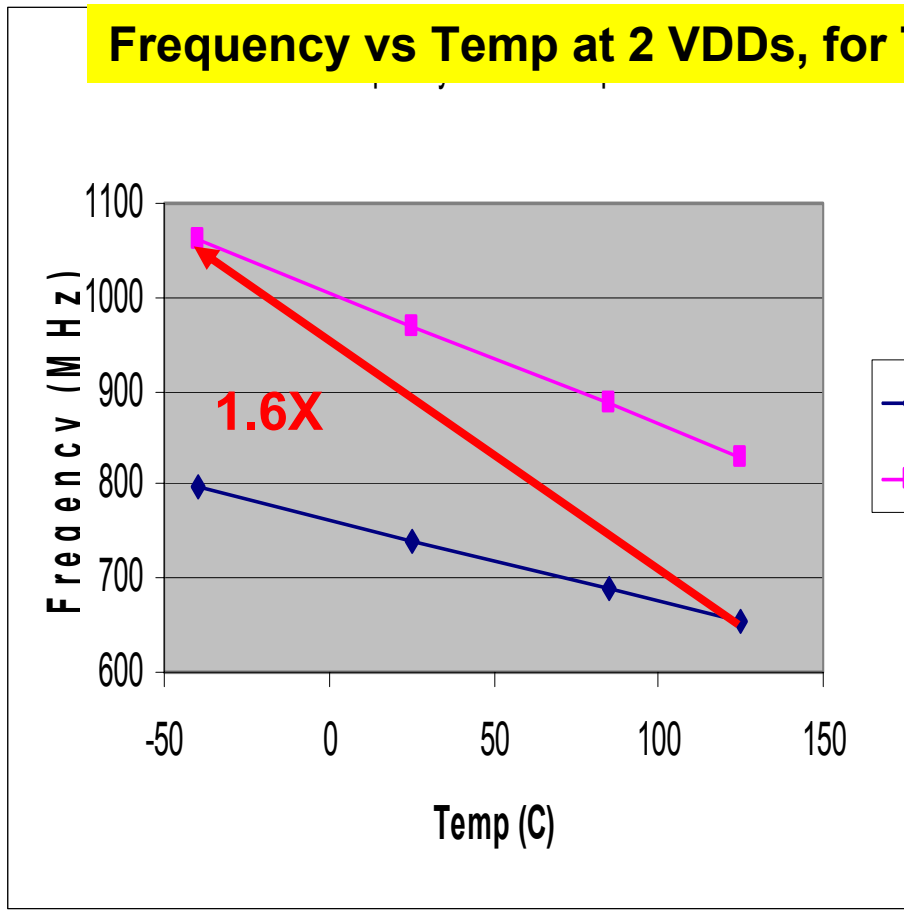
Variability of all parameters is increasing

Die Map 32nm

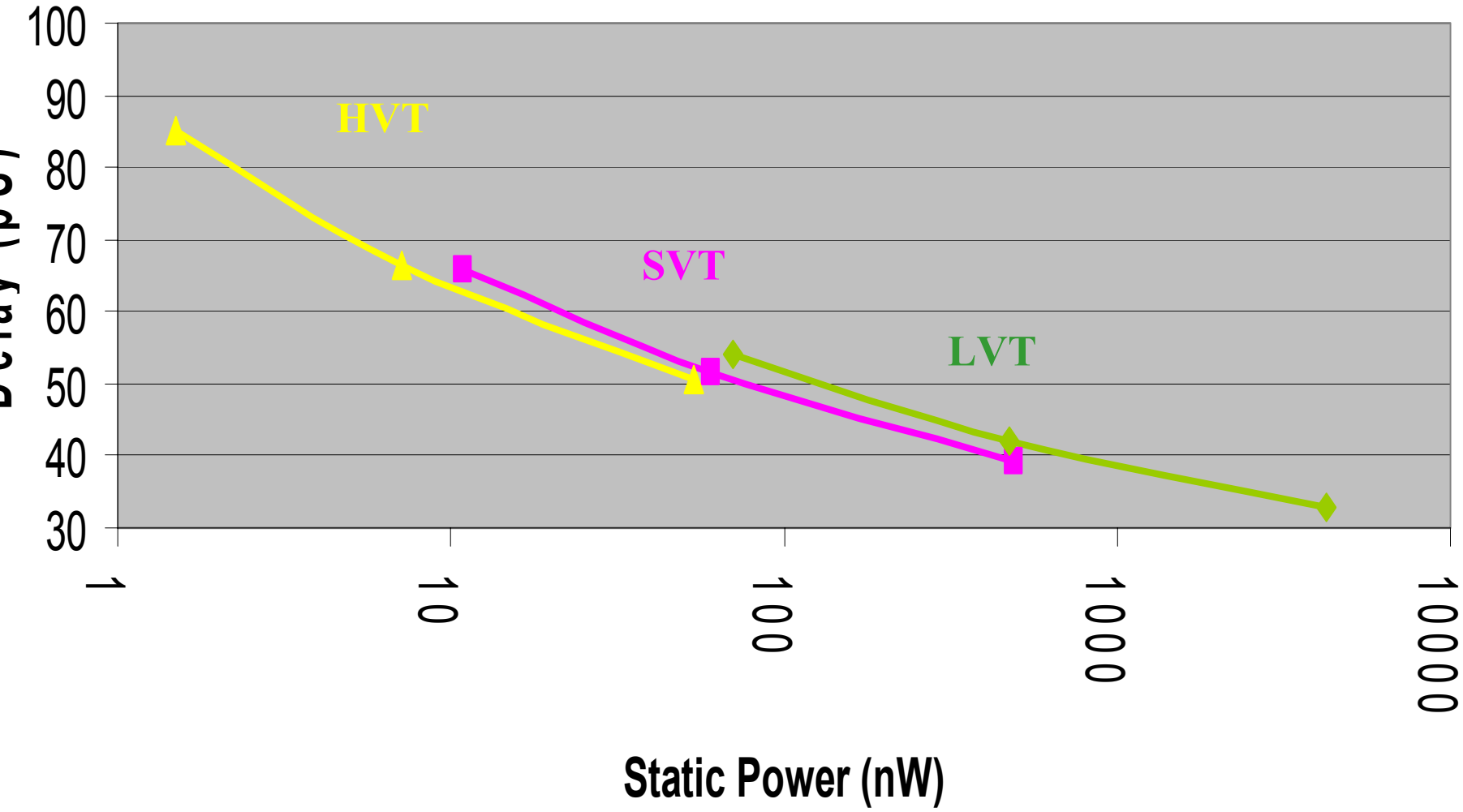


Ring Oscillator Frequency Variations (simulated in 65nm à 2 sigma)

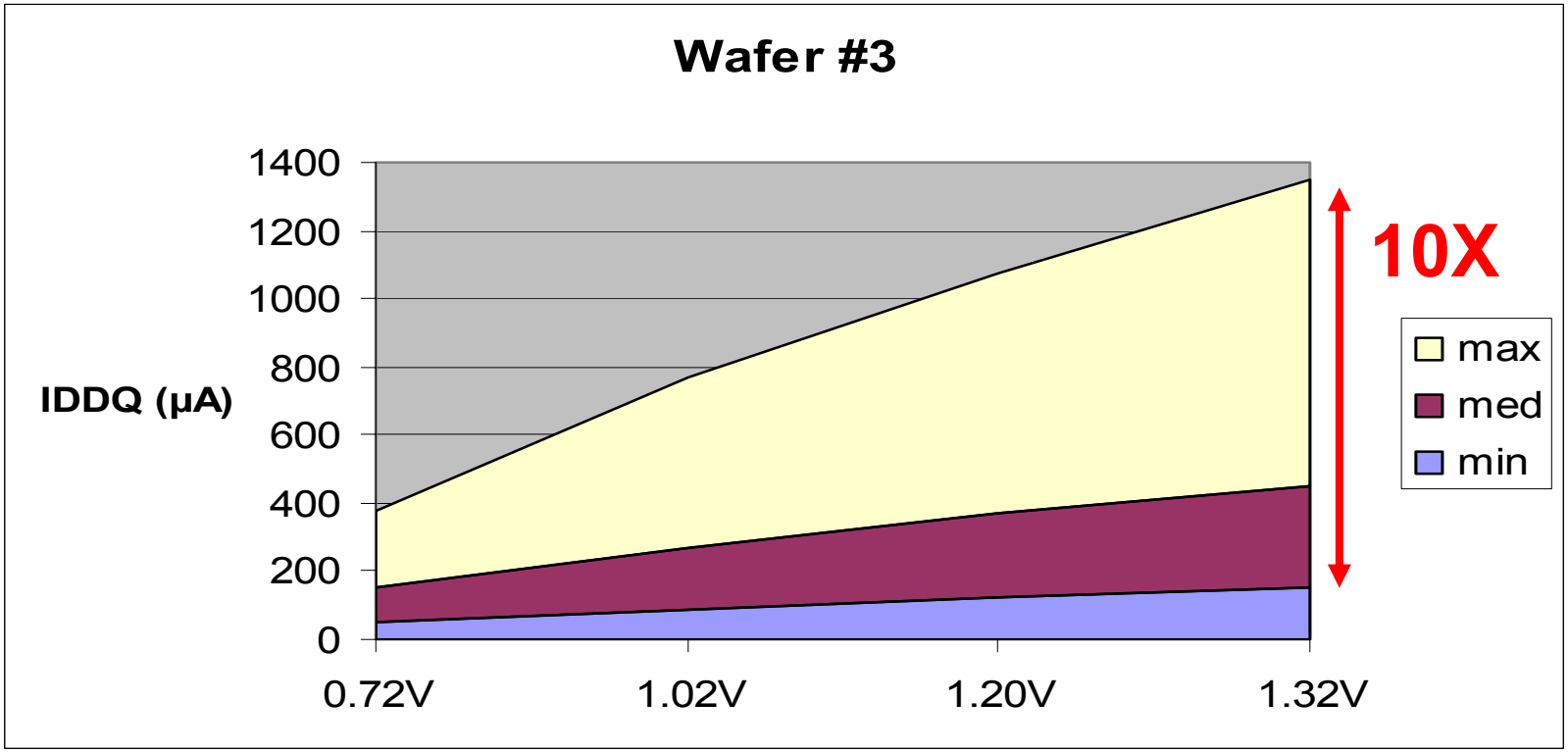




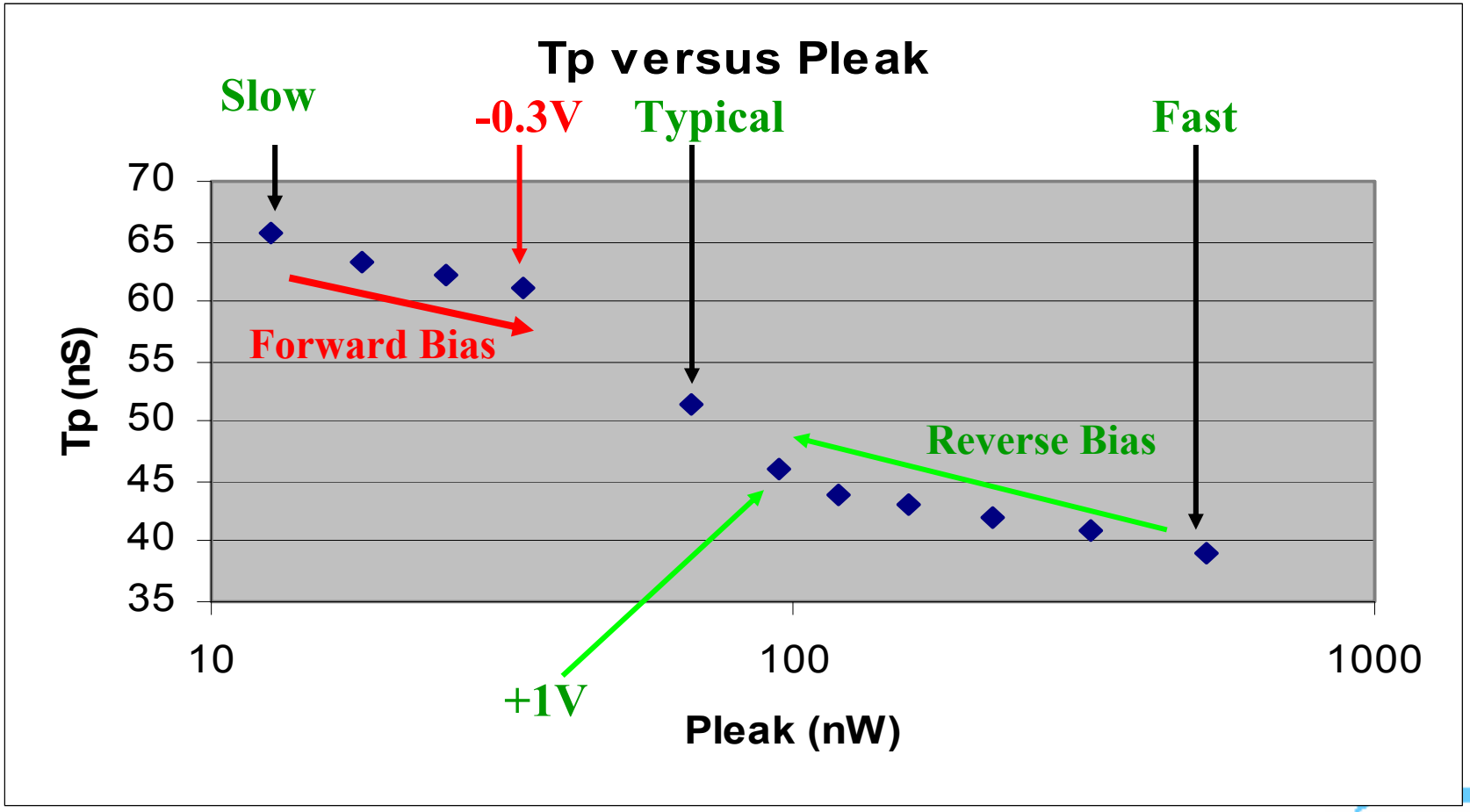
Process variations with multiple VTs



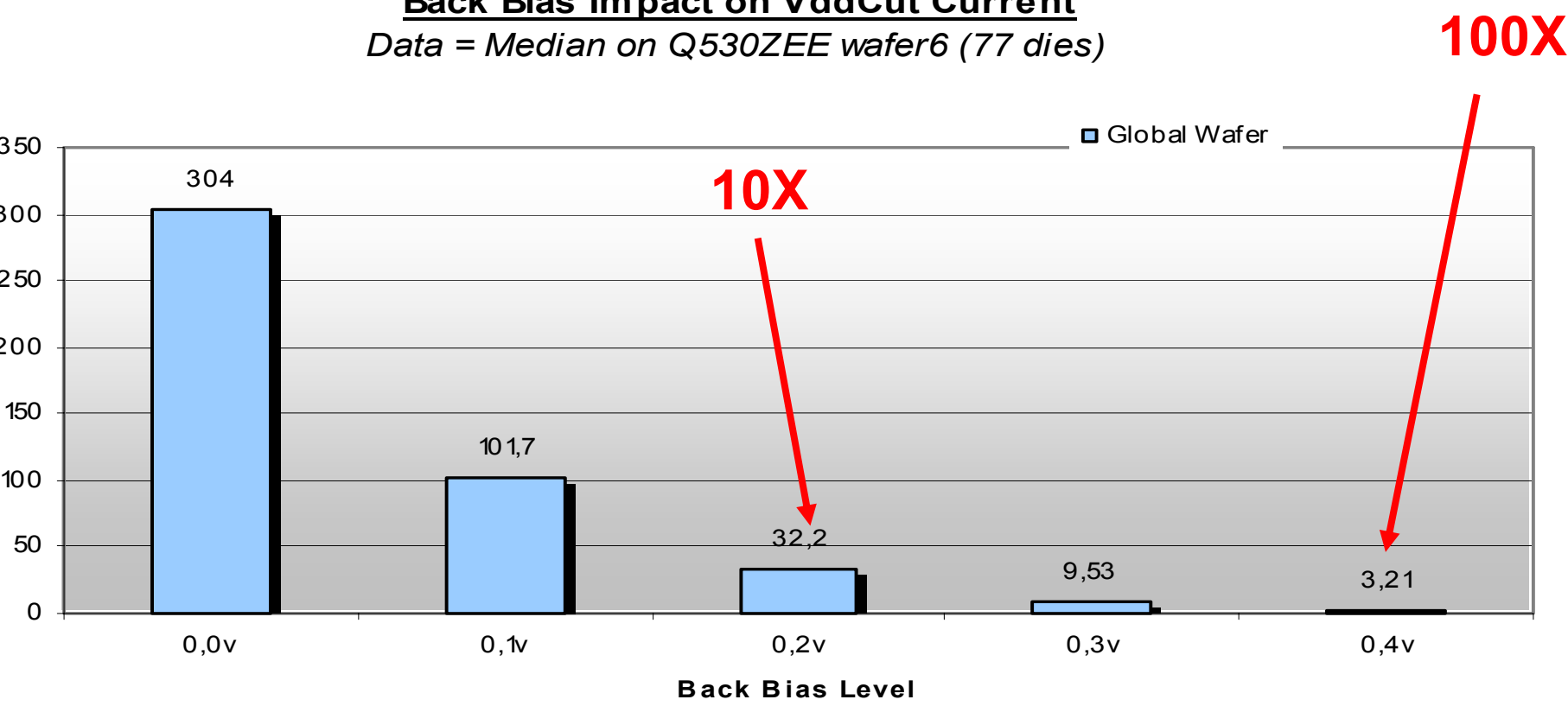
Leakage variation across wafer (a 4Mbit SRAM in 90nm)



Reduction of process spread thanks to Body Bias in 90nm:
- 10% performance gain on Slow
- 5x reduction of leakage on Fast
Less and less efficient with scaling !



Leakage reduction Technique, in Retention Mode (Source Bias in 65nm)



Vdd=1.2V **Vdd=1.0V** **Vdd=0.8V** **Vdd=0.6V** **Vdd=0.4V**

Nominal **Retention**

Defect & Variation Tolerant Architecture

- **Like in today's FPGAs, regularity and redundancy allows to sell chips with defects (regions with defects are labelled as unusable)**
- **Process spread (for both transistors and interconnects) leads to big differences between SLOW regions and FAST regions on the same chip**
- **Local Voltage and Temperature conditions also lead to big differences in « functional frequency »**
- **Cannot design anymore with a single clock:**
  - ***Either not enough performance (worst case is too SLOW to be competitive on the market, or power comsumption is too high)***
  - ***Or low yield (due to too SLOW regions)***

Locally Synchronous SRAM ?

- Today's SRAMs
  - embed a single "dummy path" for self-timing,
  - receive a unique Clock for both Read and Write (worst case FMAX: PSlow, VMin, THigh, Age=10 years NBTI stress)

- A more flexible architecture would be:
  - A big SRAM splitted into many Small SRAM blocks with a local self-timing (locally synchronous)
  - Different Read and Write cycle times
  - Externally Asynchronous: access time may vary, depending on local P-V-T-Age of the accessed block

23.6 An Asynchronous Array of Simple Processors fo DSP Applications

ISSCC 2006 by Zhivi Yu…

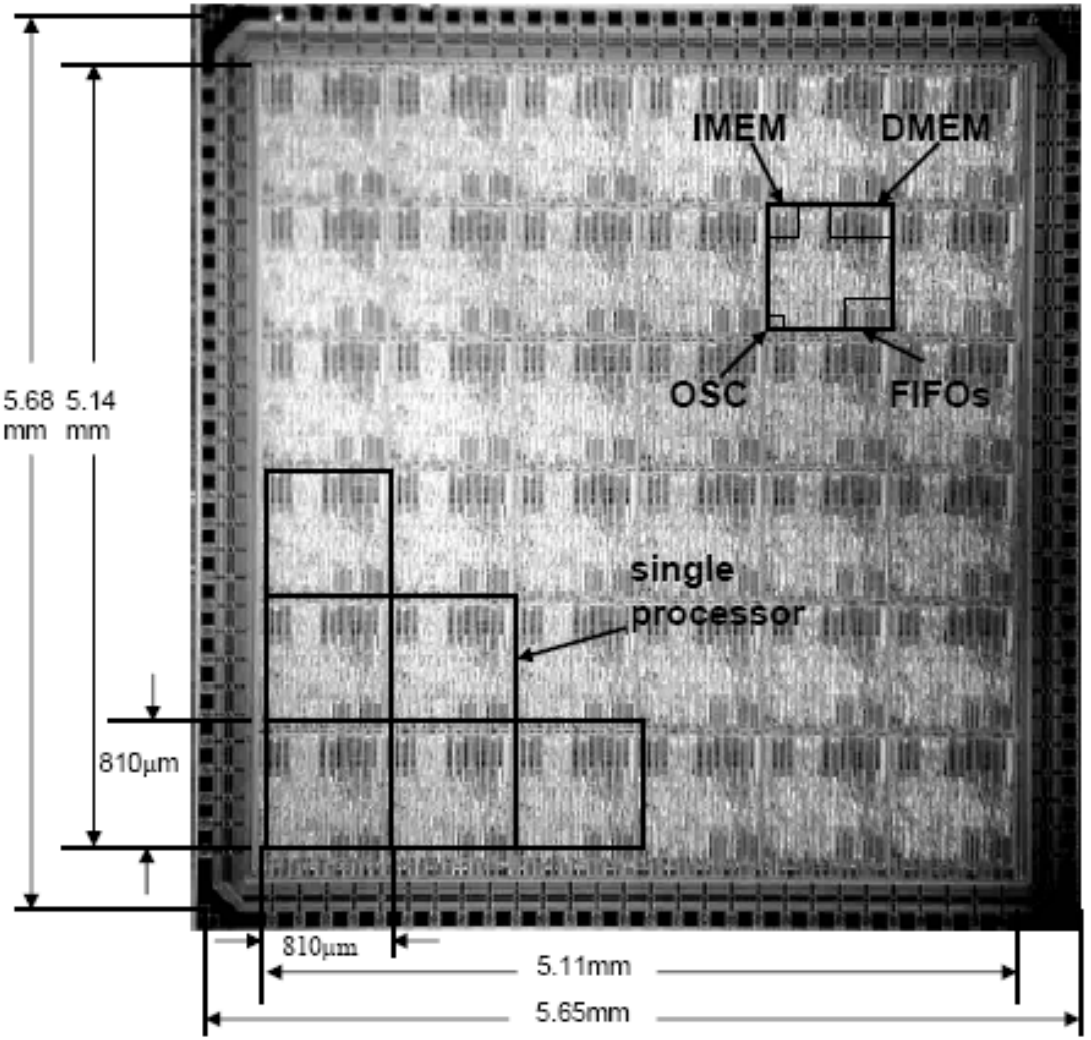
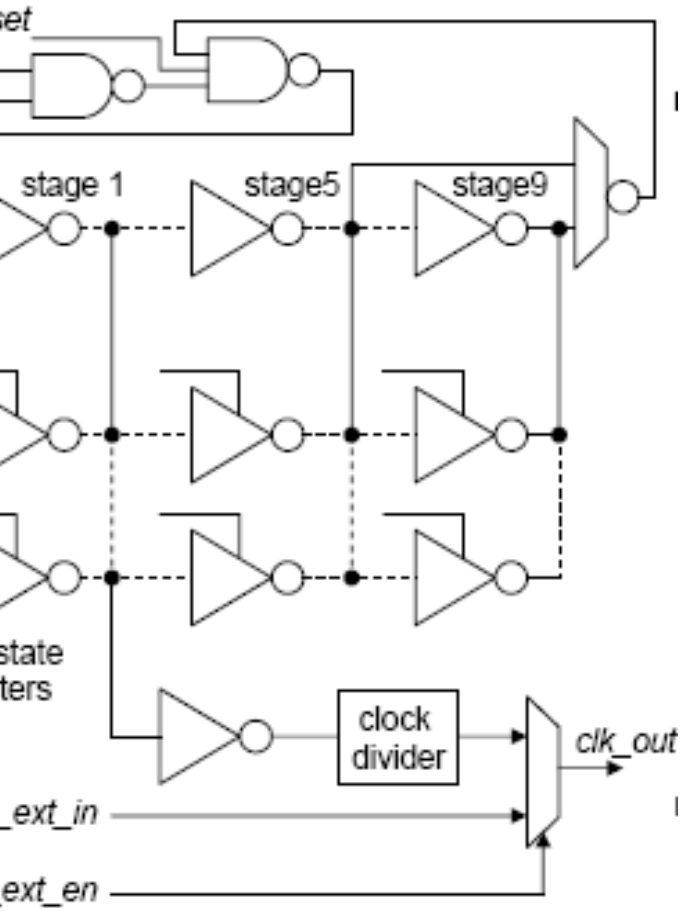


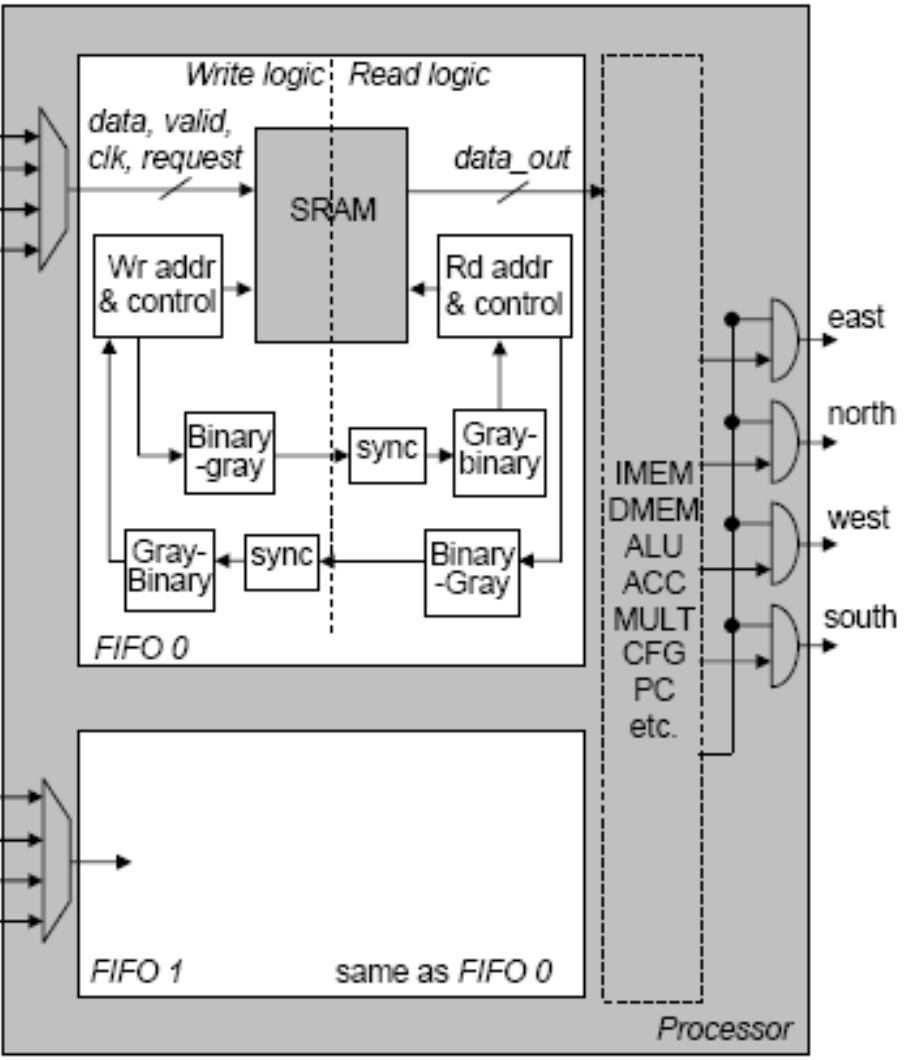
Figure 23.6.4: Programmable clock oscillator

ISSCC 2006 by Zhiyi Yu…
University of California, Davis



- Clock is generated locally
  - Exact frequency depends on local PVT
  - The programmable ring oscillators are built from standard cells and change the frequency through parallel tri-state inverters in each stage, a 5 or 9-stage configuration, and a divider that ranges from 1 to 128,
  - Processors are clocked asynchronously with respect to each other and clocking circuits permit oscillators to operate at a frequency less than the maximum
  - frequency covers a wide range from 1.66MHz to 702MHz.

Figure 23.6.4: inter-processor communication diagram, ISSCC 2006 by Zhiyi Yu… University of California, Davis



- Asynchronous communication through FIFOs
- Also applicable to any GALS or LASGA and Asynchronous NoCs

New name for this Concept: LASGA

Localy Adaptatively Synchronous and Globaly Asynchronous

SUMMARY

- **Each region (e.g. a Processor or part of it, or any block) has its own *locally generated clock, dynamically adapted to*:**
  - *local process quality (Slow/Typ/Fast)*
  - *Local voltage (F ~ 1/V)*
  - *Local temperature (F ~ 1/Temp)*
  - *Transistor ageing (F ~ 1/Age)*

- ***A region can be constrained by the Operating System***
  - *Global Power limitations (reduce F and/or VDD, or switch off)*
  - *Leakage minimization (too leaky silicon is reverse biased, or low VDD)*
  - *Slow Silicon can be boosted when more performance is needed by the O (forward bias or higher VDD)*
  - *But the ACTUAL frequency is not precisely known !*

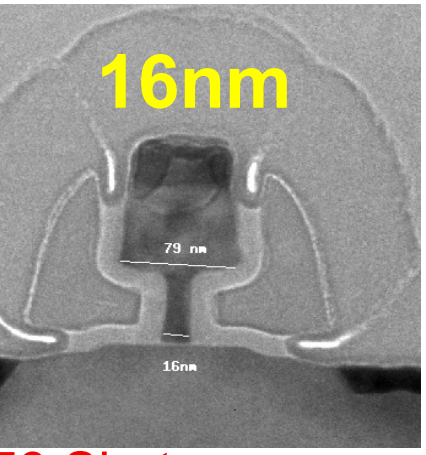
- ***Local DC/DC converter, local Process compensation may be added***

Implications of LASGA

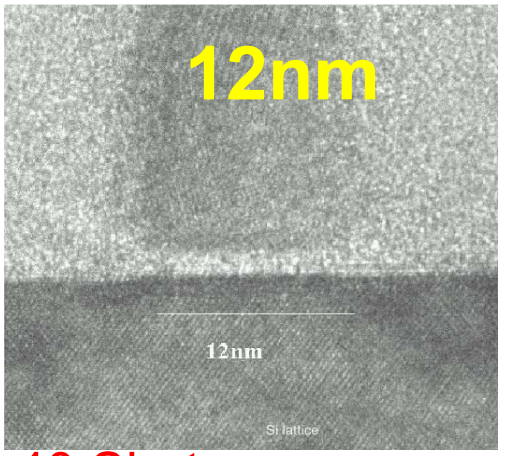
- **All communications are Asynchronous:**
  - Impact on NoC/memory architecture (no more synchronous clusters with synchronous communications)
  - Asynchronous or Serial communications (2D or 3D) :
    - Capacitive coupling between processors and NoC/Memories
    - RF links ?
- **Actual working frequency is variable*:**
  - Only MIN and MAX frequency are evaluated at design time *(completely new sign-off flow !)*
    - Simulate Clock in slowest conditions, verify that design is working at that frequency FMIN
    - Simulate Clock in fastest conditions, verify that design is working at that frequency FMAX
    - What happens in the middle ??? *(design must be robust !)*
  - Task duration is variable:
    - *: OS may know the current frequency (counter + divider + real time clock reference)
    - OS can react to ask for more or for less performance, by adjusting VDD for a block (F is automatically adjusted): V-controled DVFS !

DISPERSIONS-VARIABILITY:

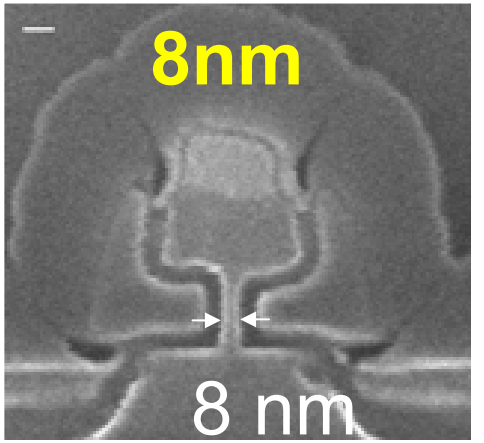
DISCRETENESS OF MATTER (CROLLES EXPERIMENTAL MOSFET SAMPLES):



53 Si-atoms
& 3.5 B-dopants
1e19cm-3) under gate



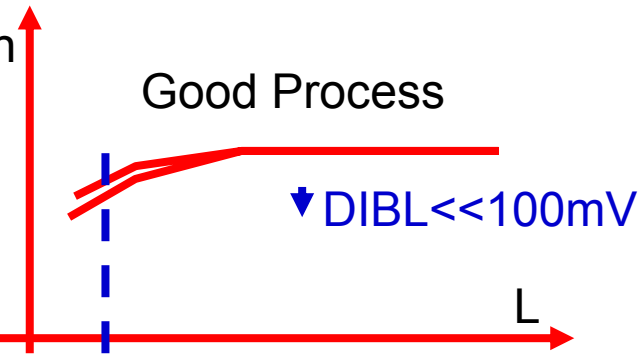
40 Si-atoms
& 2.6 B-dopants
(1e19cm-3) under gate

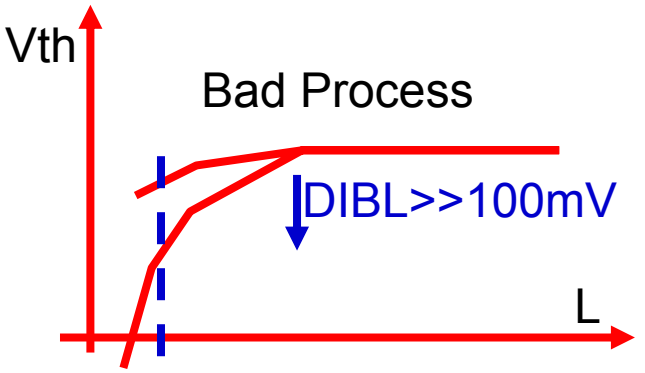


27 Si-atoms
1.7 B-dopants
(1e19cm-3) under gate

Courtesy: Thomas Skotnicki, ST Crolles

PROCESS DISPERSIONS:





Part2
Challenge for the future:

Improve MOPS/Watt @ constant dynamic power

Pat Gelsinger
Senior Vice President & CTO
Intel Corporation

June 8, 2004

Transistors (and silicon) are free

Power is the only real limiter

Optimizing for frequency AND/OR area may achieve neither

llenges seen by Bernard Meyerson, chief technologist for IBM Systems & Technology Group

e Sum Is Greater Than the Whole", talk published on March 5, 2006 in Electronic News

(CELL) is designed to have tremendous processing capabilities through a *multitude of accelerators working in concert with a core* processor.

***tradeoffs to balance numerous aspects of performance* --clock frequency, powe utilization, actual processing or data throughput, integration of the processor elements with the rest of the system in terms of communication buses, the bus architecture, the software…**

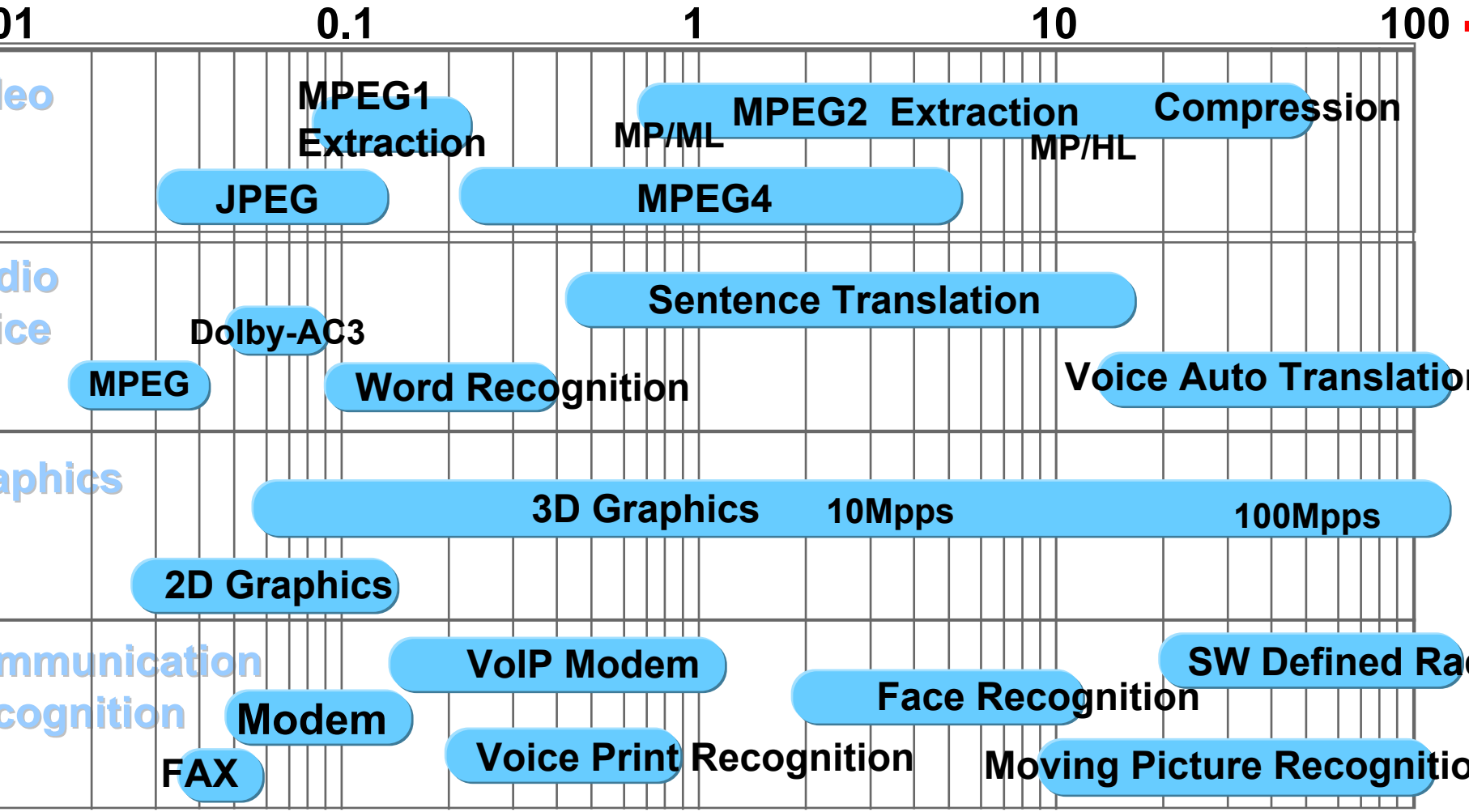
***Innovation will be the driver of performance, rather than scaling*.**

Going forward, you can say Moore's Law continues because you will continue t make each generation smaller. However, … *further shrinking of the chip does not ensure higher performance*. That's the key.

if you run a *processor frequency at half* of what it's capable of, you could conceivably *save five times the amount of power*.

Required Performance for Multi-Media Processing is increasing forever (yet another "law" ?)

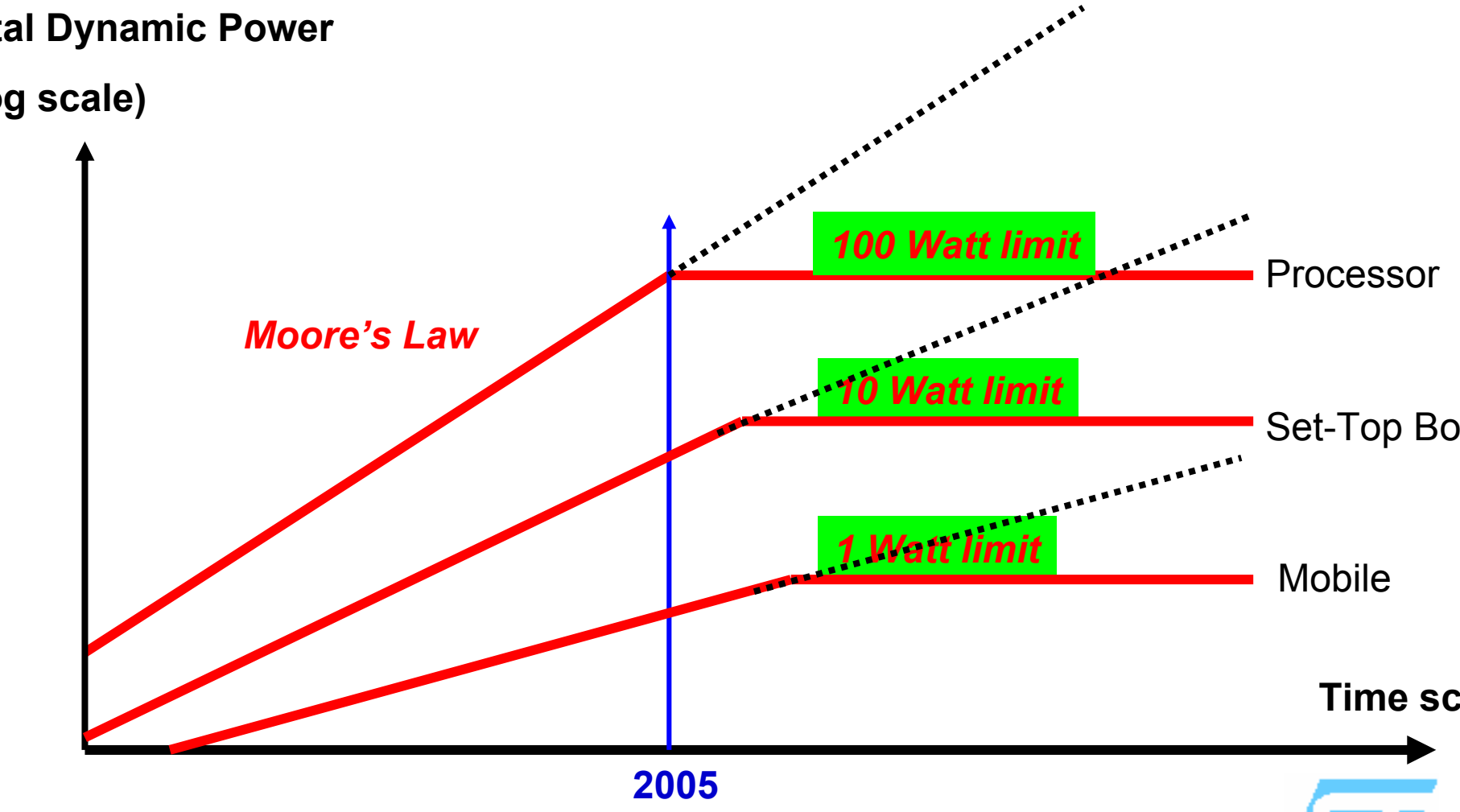
(source ITRS Design ITWG July 2003)



GOPS = Giga Operations Per Second

Dynamic Power is THE limiting factor

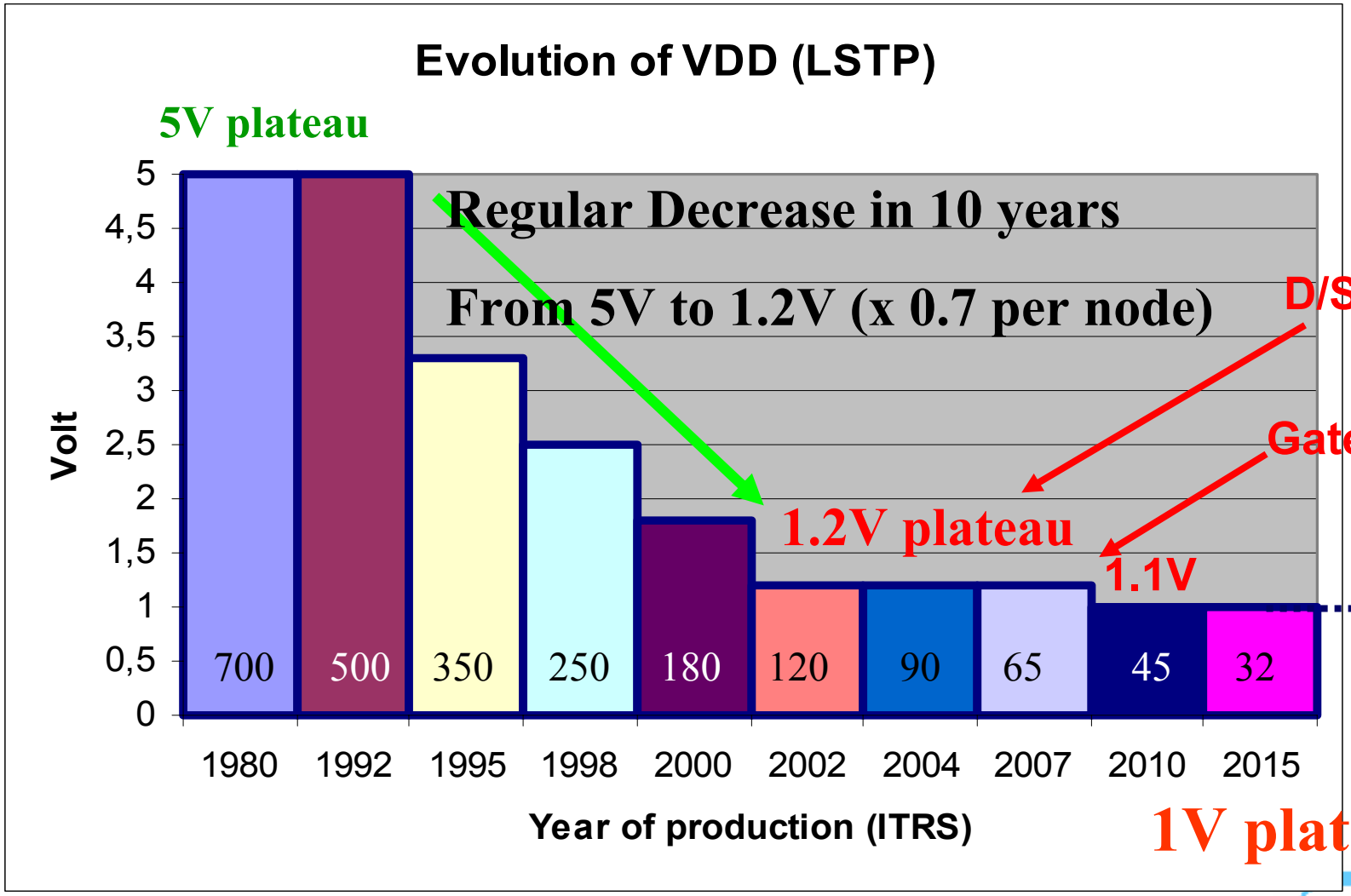
(max value is application dependant)



Silicon is free, Power is limiting

- Suppose we have today a 50 mm2 chip consuming 5 Watt, containing 9 small Processing Elements (Micro, or Core, or DSP, or FSM, …)
- Scaling will allow to implement ~200 Processing Elements on 50 mm2 at node 22nm
- *If we follow the ITRS*, processing power would be tremendously improved (150X), but Power Comsumption would jump to 100 Watt !

VDD (no more) scaling can explain the « power crisis »



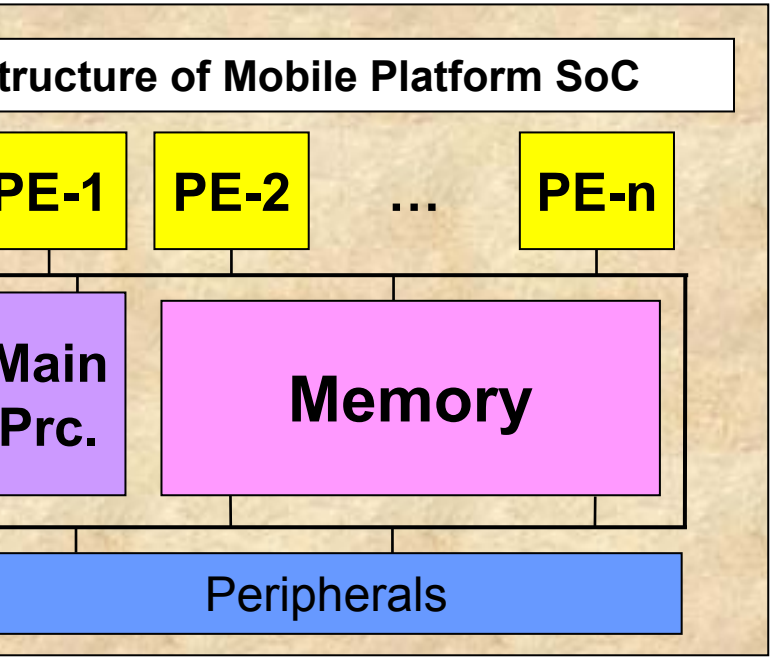
Definition of “Mobile Platform SoC”
om STRJ: Semiconductor Roadmap for Semiconductor Japan)

tructure of Mobile Platform SoC

- Main Processor + Processing Engines (PE) + Peripherals.

rocessing Engine (PE)

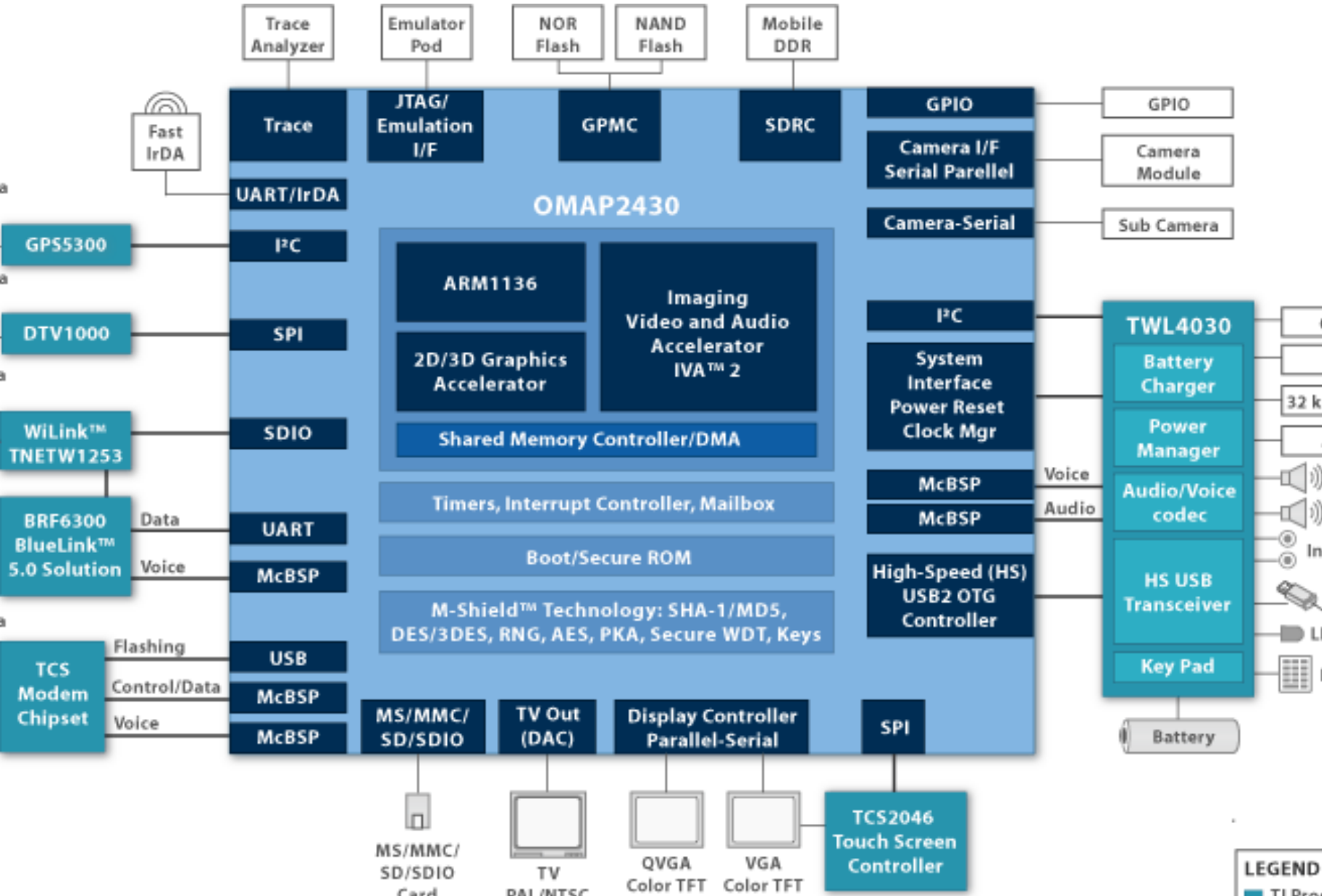
- PE is a processor customized for a specific function.
- A complex function is implemented as a set of PEs.



- **Die Size:**
  - **50 mm² (Constant)**
- **Main Processor**
  - **500kG Logic (Constant)**
  - **512k bit Memory (Constant)**
- **Processing Engine**
  - **150kG Logic (Constant)**
  - **64k bit Memory (Constant)**
  - **Clock Frequency:**
    proportional to the device performance
- **Memory:**
  - **1,024k bit per PE**
- **Peripherals: 1MG (Constant)**

OMAP 2430

From TI website



ATRS: an Alternative Technology Roadmap for Semiconductors

- Suppose we have today a 50 mm2 chip consuming 5 Watt, containing 9 small Processing Elements (Micro, or Core, or DSP, or FSM, …) *(same chip as on previous slide)*
- Scaling allows to implement ~200 Processing Elements on 50 mm2 at node 22nm
- If we follow the ATRS, processing power would be modestly improved (5.7X = +16.5%/year), but Power Comsumption would be stable at 5 Watt:
  - with “only” 100 Processing Elements @ frequency = 200 MHz
  - Or 200 Processing Elements @ frequency 100 MHz

ATRS: an Alternative Technology Roadmap for Semiconductors

- So, what is the secret of ATRS to improve MOPS/Watt @ constant Power ?

- ATRS is proposing:
  - a reduction of the Frequency of operation
  - A “reasonnable” reduction of Voltage

- This is relaxing the pressure on Technologists:
  - Improve mobility, but at low voltage
  - Control leakage: both short channel and gate leakage

- ATRS is requesting innovation in multi-processor architectures, for a good usage of all that “free silicon area”

ATRS: a possible SCENARIO

Keeping Total Dynamic Power CONSTANT with +16.5% / year Processing Power Increase

		2004	2007	2010	2013	2016
nology Node (nm)		90	65	45	32	22
(V)		1,20	1,08	0,97	0,87	0,79
uency (GHz)		0,40	0,34	0,29	0,25	0,21
# of PE due to fixed area		9	21	47	98	208
# of PE due to fixed power		9,00	16,30	29,50	54,00	**98,00**
amic Power (W)		5,18	5,17	5,16	5,20	5,20
essing Power (GOPS)	normalized to 2004	1,00	1,54	2,37	3,68	5,68
essing Power increase			1,54	1,54	1,56	1,54
ce Performance	normalized to 2004	1	0,80	0,64	0,51	0,41
OS IDSAT - ATRS	µA/µM	440	396	356	321	289
OS IDSAT – ITRS 2003	µA/µM	440	510	760	880	990

Scali

facto

0.9

0.8

1.0

0.9

1.2

IDSAT/µm = -10%/node instead of +20%/node

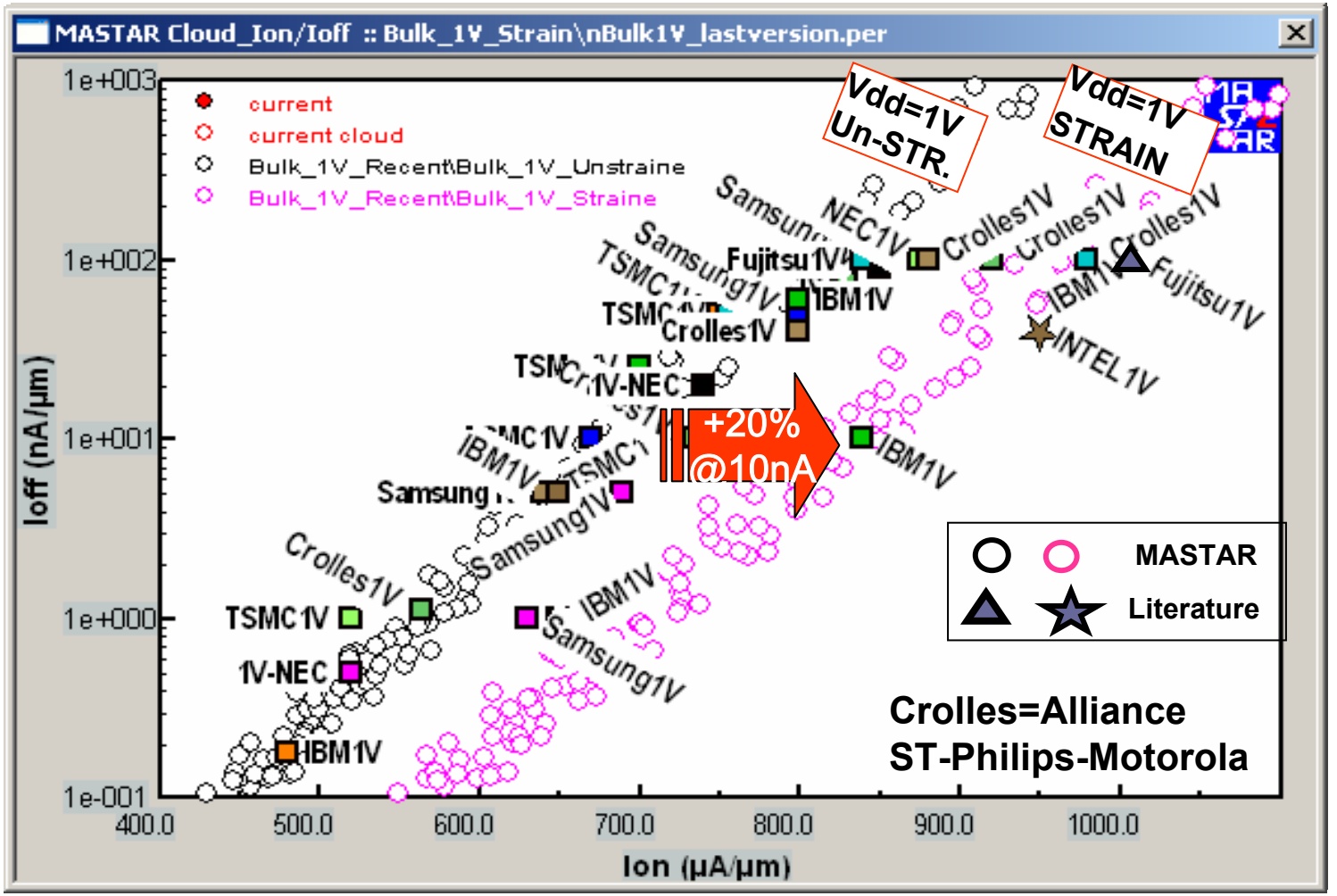
Challenge:
Improve MOPS/Watt @ constant dynamic powe
SUMMARY

- Approach #1 **(ATRS proposal)** :
  - **massive parallelism (area is free)**
  - **+ Reduce Frequency + Innovative Architecture**
  - **=> Relax device performances**

- Approach #2:
  - **Improve device performances (better mobility)**
    - **HP-SOI, GeOI, Double Gate (FinFET, SON, …)**
    - **HighK for Low Gate Leakage, UltraLowK for Interconnects**
  - **+ Reduce Voltage**

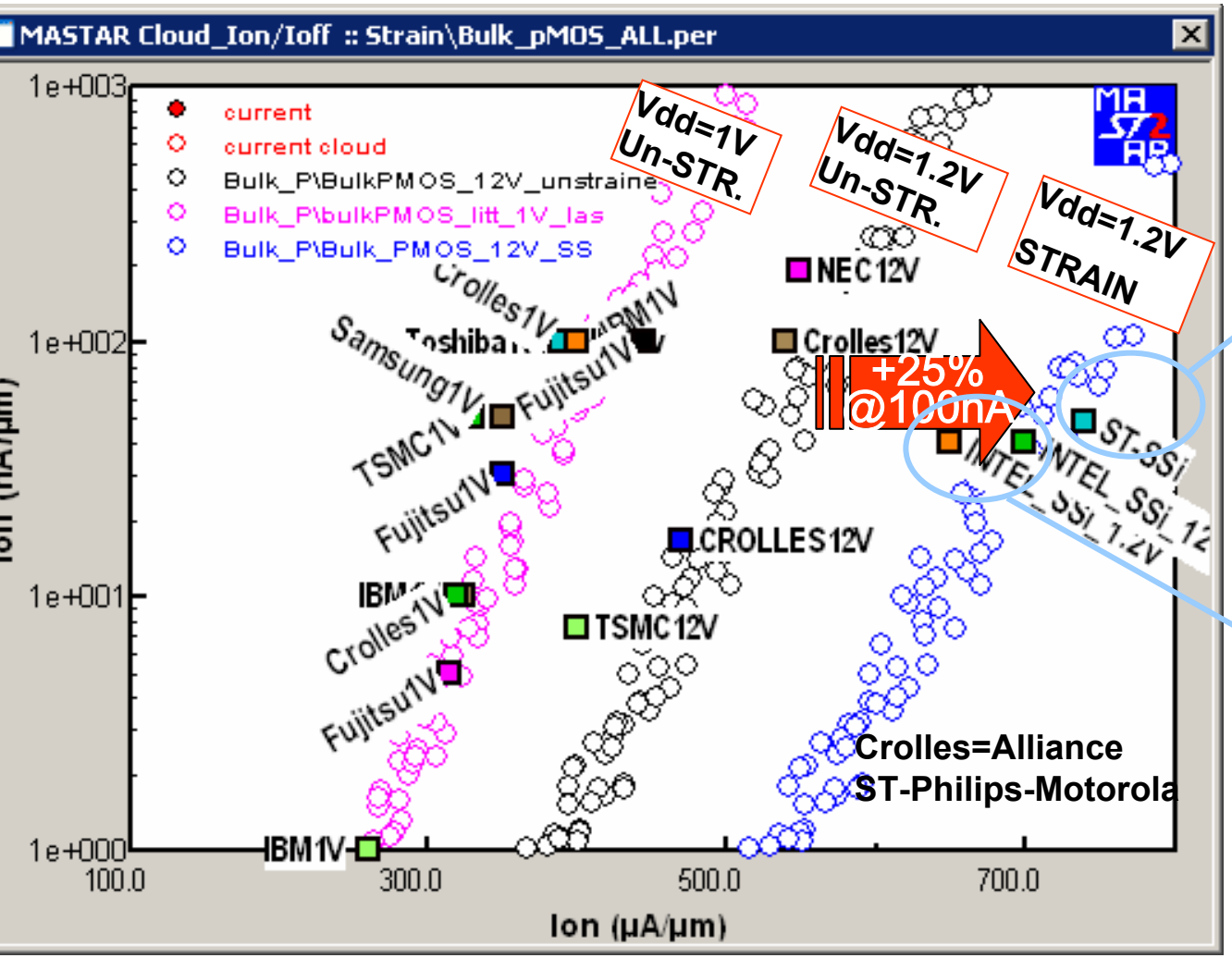
- Approach #3: a mix of #1 and #2 is the most probable ?

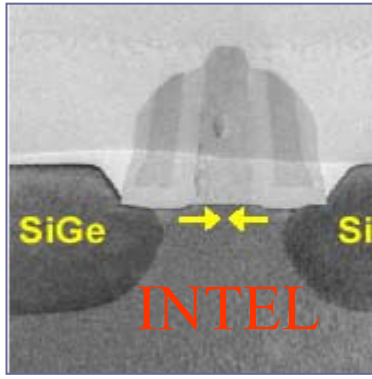
STRAINED-Si-NMOS : State-of-the-ART



Courtesy: Thomas Skotnicki, ST Crolles

STRAINED-Si-PMOS : State-of-the-ART





Courtesy: Thomas Skotnicki, ST Crolles

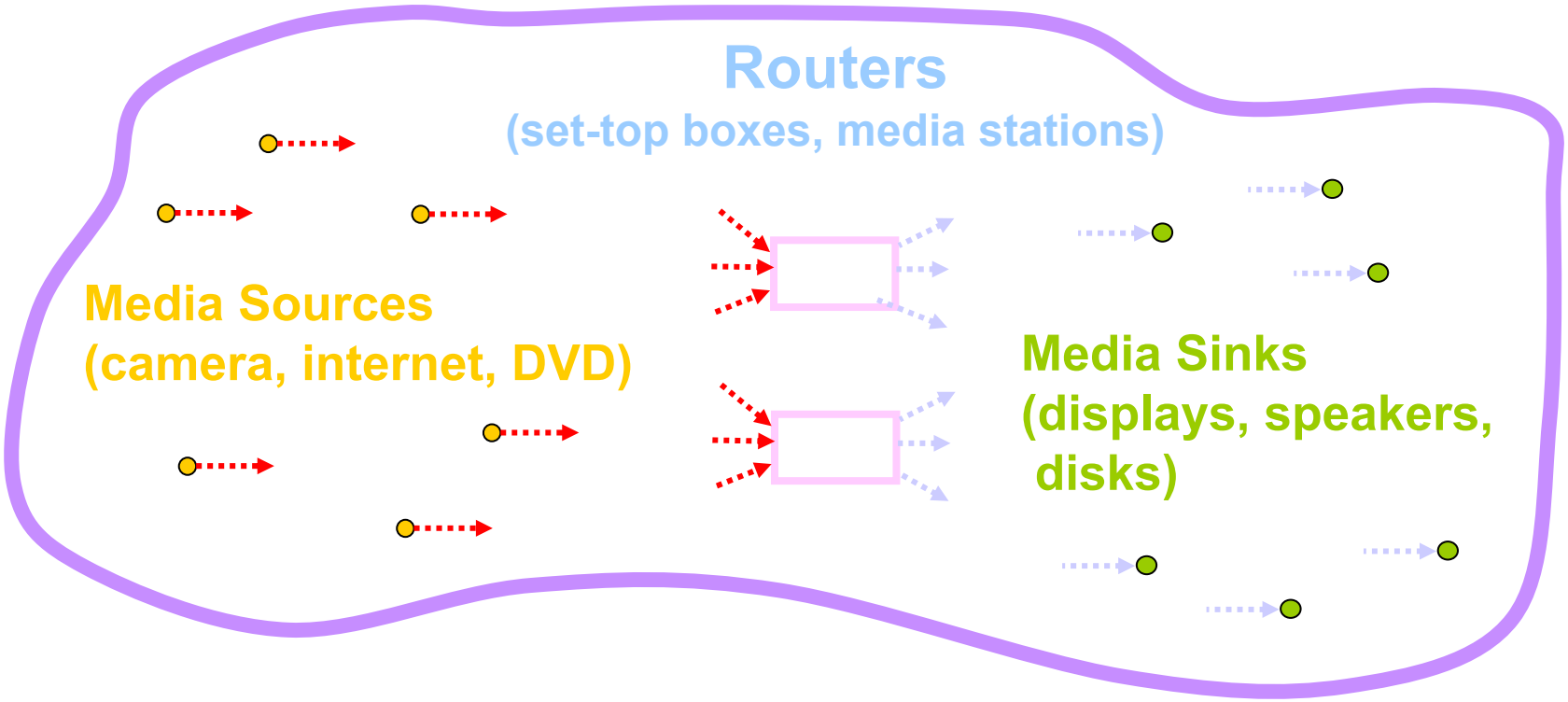
Future of communications

- Communications between real world and processing units is Asynchronous by nature
- Intra-chip communications will be more and more Asynchronous

ealing with the Myriad of Protocols and Format

Put the Intelligence in the Network: "The Ambient Home Router"

Jan M. Rabaey, BEARS 2

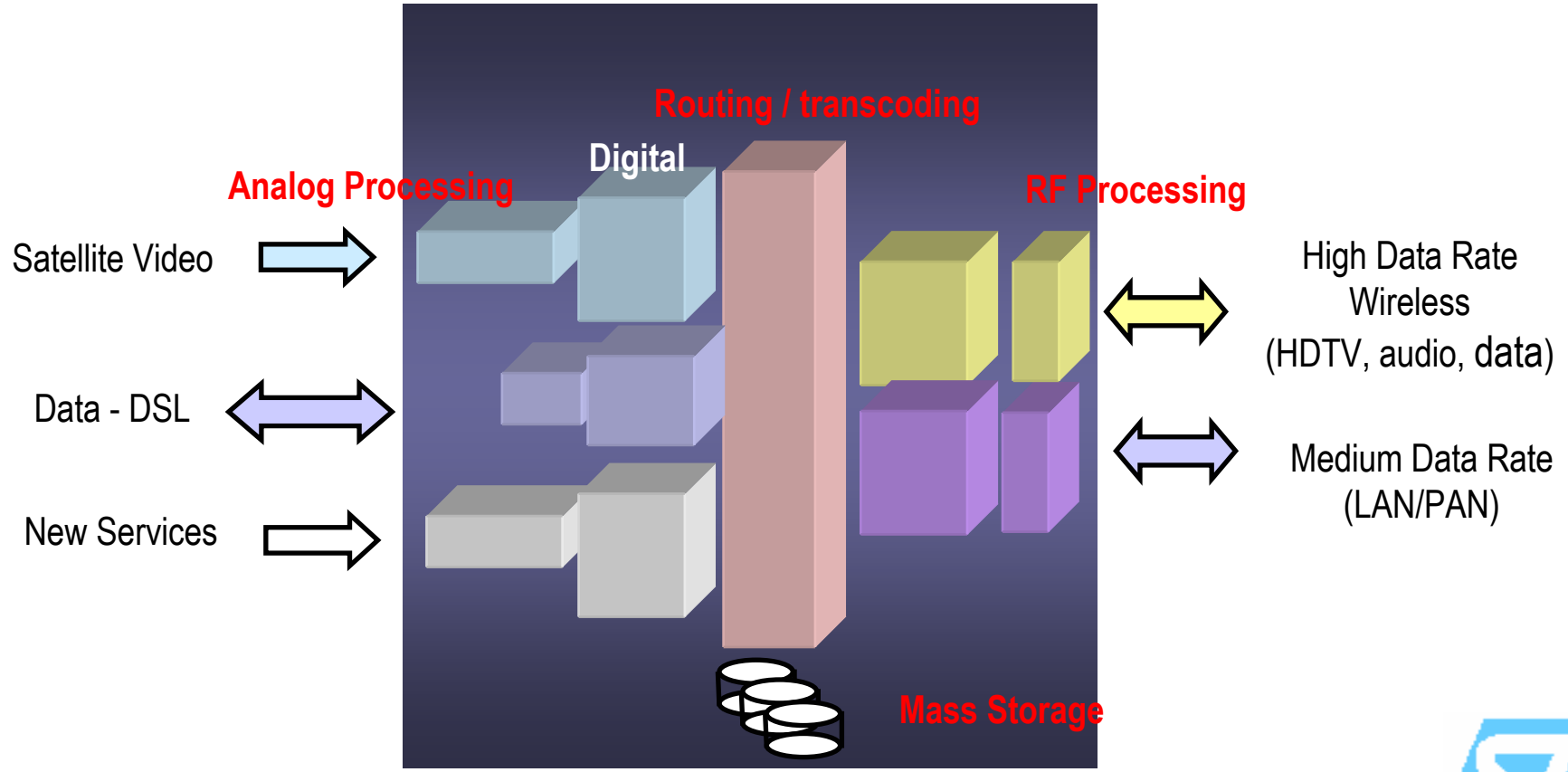


Home routers: Provide on-the-fly protocol conversion and trans-coding based on properties of source and destination devices

"Seamlessly connect everything to anything – or do even better than th

The Ambient Home Router Implementation Challenge

Must process multiple real-time high-data rate streams from physical interface through protocol stack and signal processing (TOPS) in fully programmable and upgradable fashion at a extremely constrained cost and power budget.



CONCLUSION

- Both approaches:
  - LASGA to handle increasing process variations
  - ATRS to maintain Dynamic Power at an acceptable level

- Require to put in place efficient Asynchronous communications between hundreds of Processing Elements (which may run in a Locally Synchronous way or any other efficient way
  like fully Asynchronous ???)

THANK YOU

Jean-Pierre.Schoellkopf@ST.com