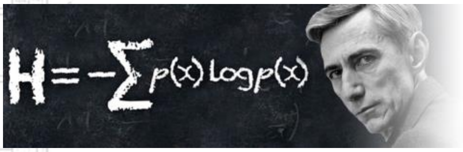
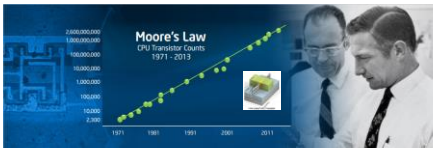
High Throughput Coding Workshop
DLR, February 2019

High Throughput Coding
An Implementation Centric View

...when spectral efficiency meets nm/pJ...
...when Shannon meets Moore...

Norbert Wehn





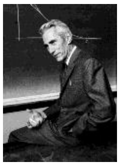
DFG Deutsche Forschungsgemeinschaft

EPIC
EPIC Grant Agreement No. 760150

Technische Universität Kaiserslautern

Birth of Information Theory 70 Years Ago

Reprinted with corrections from *The Bell System Technical Journal,*
Vol. 27, pp. 379–423, 623–656, July, October, 1948.



A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

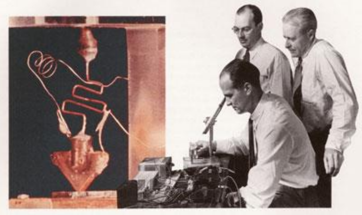
THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

Theorem 11: Let a discrete channel have the capacity C and a discrete source the entropy per second H. If $H \leq C$ there exists a coding system such that the output of the source can be transmitted over the channel with an arbitrarily small frequency of errors (or an arbitrarily small equivocation). If $H > C$ it is possible to encode the source so that the equivocation is less than $H - C + \epsilon$ where ϵ is arbitrarily small. There is no method of encoding which gives an equivocation less than $H - C$.

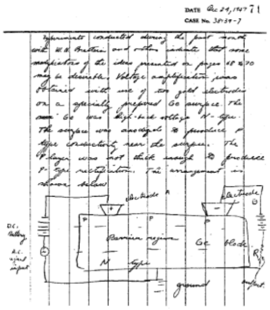
Birth of Microelectronics 70 Years Ago

Christmas 1947 (Bell): Bardeen, Brattain discover point contact transistor

January 1948 (Bell): Shockley discovers junction transistor



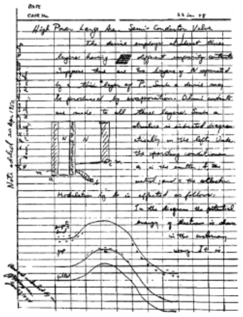
2 gold foils pressed onto germanium
- Gold contact in forward direction
- Gold contact in backward direction



DATE Dec 24, 1947 71
CASE No. 38139-7

Experiments conducted during the past month with W. H. Brattain and others indicate that some modifications of the ideas presented on pages 68 to 70 may be desirable. Voltage amplification was obtained with use of two gold electrodes on a specially prepared Ge surface. The Ge was high-back voltage N-type. The surface was anodized to produce P type conductivity near the surface. The P-layer was not thick enough to produce P-type rectification. The arrangement is shown below

FIG. 3. Entry in Bardeen's lab notebook dated 24 December 1947, giving his conception of how the point-contact transistor functions. (Reprinted by permission of AT&T Archives.)



DATE
CASE No.
23 Jan 48

High Power Large Area Semi-Conductor Valve

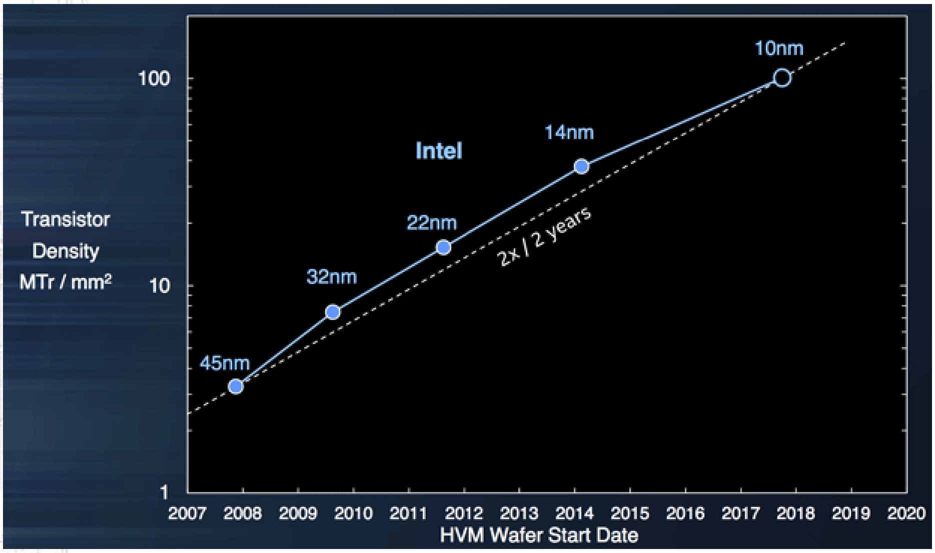
The device employs at least three layers having different impurity contents. Suppose there are two layers of N separated by a thin layer of P. Such a device may be produced by evaporation. Ohmic contacts are made to all three layers. Such a structure is indicated diagrammatically on the left. Under the operating conditions a is the emitter, b the control, and c the collector. Modulation by b is effected as follows: In the diagram the potential energy of electrons is shown in the customary way. It is

Note added 20 Apr 1950

FIG. 4. Entry in Shockley's lab notebook dated 23 January 1948 recording his conception of the junction transistor. He wrote this page at home on a piece of paper, which he later pasted into his notebook. (Reprinted by permission of AT&T Archives.)

Bell System Tech. Journal, Vol 27, 1948: A Mathematical Theory of Communication, C. Shannon
Bell System Tech. Journal, Vol 28, 1949: The Theory of p-n junctions in semiconductors and p-n junctions transistors, W. Schockley

Moore’s Law Forever ?



Source: G. Intel

Limits of Moore’s Law

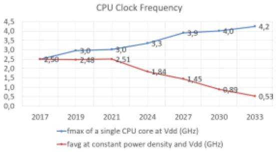
Cost

- Wafer cost 28nm -> 7nm: more than **doubles**, but area density increases by **6x**
- Average IC design cost for 16nm/14nm chip is ~ $80 million
- In 7nm it will cost > $200 million

⇒ Only high volume chips justify use of advanced technology nodes

Performance gain

- 28nm -> 7nm: **3x** improvement in frequency (4 nodes)
- 2018 -> 2033: **1.7x** frequency gain (7 nodes)



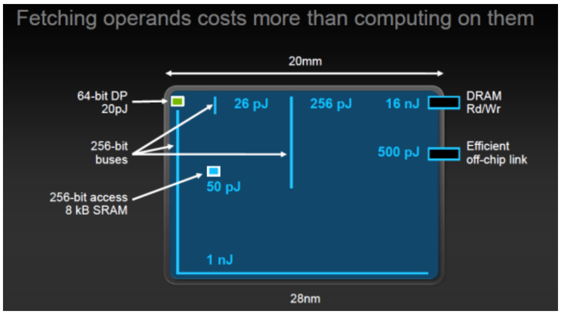
Power density

- Power per transistor decreases slower than transistors density increases
  ➔ power per mm^2 increases
- Until 2033, power density increases by **8x**
- Same TDP: frequency has to be reduced by **8x** 4.2GHz -> 0.5GHz

Limits of Moore’s Law

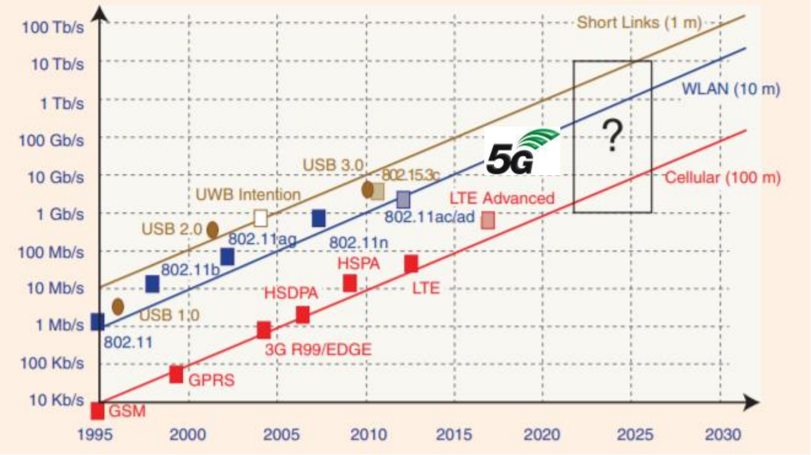
Interconnect

- 14nm technology delay for 1mm wire ~400ps
- Until 2033
  - 9% delay **improvement** in logic per technology node **without wires**
  - 10% node-to-node **penalty** for data path **with tight metal pitches**
- Energy challenge


Fetching operands costs more than computing on them

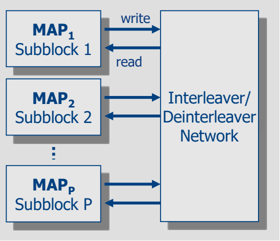
Source: NVIDIA

Wireless Communication



Source: G. Fettweis

Microelectronic Contribution to Channel Coding

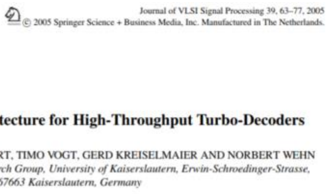

MAP_1 Subblock 1
write
read
MAP_2 Subblock 2
:
MAP_P Subblock P
Interleaver/ Deinterleaver Network

2 Turbo-Code decoders in different technologies

- Both decoders designed with the same methodology
- Similar basic architecture: exploit spatial parallelism, sub-blocks on several MAP decoders in parallel

Decoder 1 (2004)

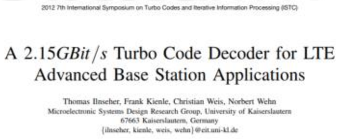
- UMTS compliant decoder in **180nm** technology
- Max frequency **166 MHz**
- **16** MAP decoders in parallel
- Throughput **80 Mbit/s** @ 6 iterations
- **30 mm^2**


Journal of VLSI Signal Processing 39, 63–77, 2005
© 2005 Springer Science + Business Media, Inc. Manufactured in The Netherlands.

A Scalable System Architecture for High-Throughput Turbo-Decoders

MICHAEL J. THUL, FRANK GILBERT, TIMO VOGT, GERD KREISELMAIER AND NORBERT WEHN
Microelectronic System Design Research Group, University of Kaiserslautern, Erwin-Schroedinger-Strasse, 67663 Kaiserslautern, Germany

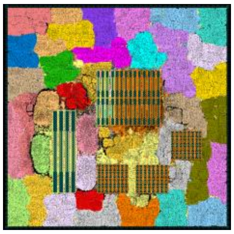
Channel Coding



2012 7th International Symposium on Turbo Codes and Iterative Information Processing (ISTC)

A 2.15*GBit/s* Turbo Code Decoder for LTE
Advanced Base Station Applications

Thomas Ilnseher, Frank Kienle, Christian Weis, Norbert Wehn
Microelectronic Systems Design Research Group, University of Kaiserslautern
67663 Kaiserslautern, Germany
{ilnseher, kienle, weis, wehn}@eit.uni-kl.de



Decoder 2 (2011)

- LTE compliant decoder, **65nm** technology
- Max frequency **450 MHz**
- **32** parallel MAP decoders
- Throughput **2.15Gbit/s** @ 6 iteration
- **7.7 mm²**

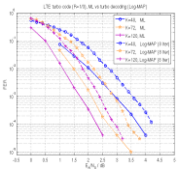
Comparison

- **180nm**, 130nm, 90nm, **65nm**
- Throughput increase **27x**, but frequency increase only **3x**
- Improvement in area efficiency (area/throughput) **100x**

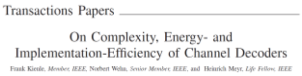
⇒ Progress due to microelectronic mainly in area efficiency

⇒ Throughput increase mainly due to code, algorithm, architecture: e.g. conflict free interleaver, NII, radix-4, re-computation, advanced normalization, larger parallelism…

Communications Performance versus Implementation Efficiency



⟷



Transactions Papers

On Complexity, Energy- and
Implementation-Efficiency of Channel Decoders

Frank Kienle, *Member, IEEE,* Norbert Wehn, *Senior Member, IEEE,* and Heinrich Meyr, *Life Fellow, IEEE*

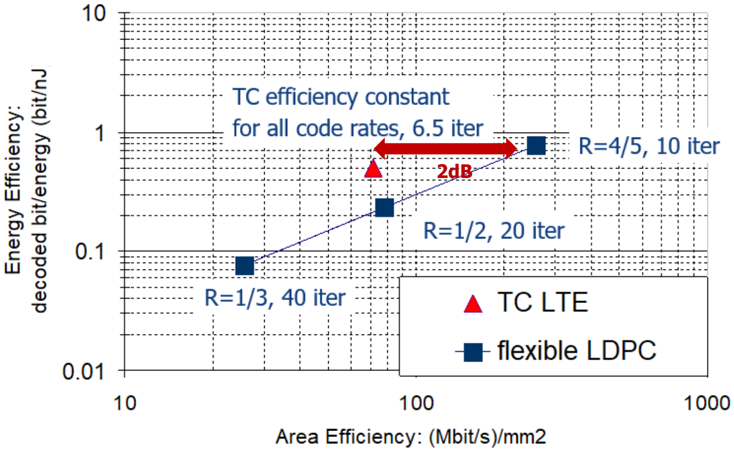
Various decoders in same technology, same design methodology

- Communications performance: FER/BER over SNR
- Implementation efficiency
  - Area efficiency: decoded bits/s/mm²
  - Energy efficiency: decoded bits/s/power = decoded bits/energy

Comparison of TC decoder and LDPC decoder

- LTE compliant Turbo-Code decoder: rate 1/3…9/10 (puncturing)
- Flexible LDPC decoder: rate 1/4…9/10
- Blocksize for comparison 6154 bits

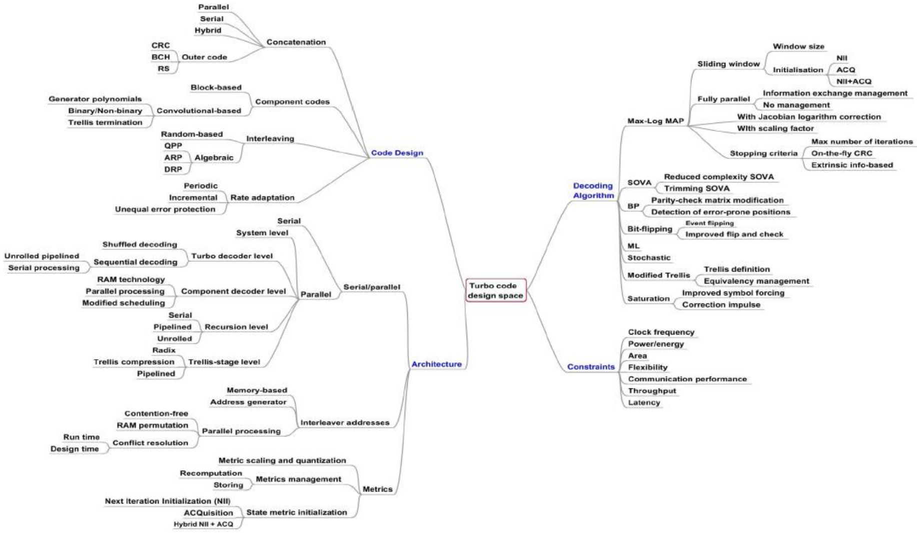
Communications Performance versus Implementation Efficiency



Strong interrelation communication performance and implementation efficiency

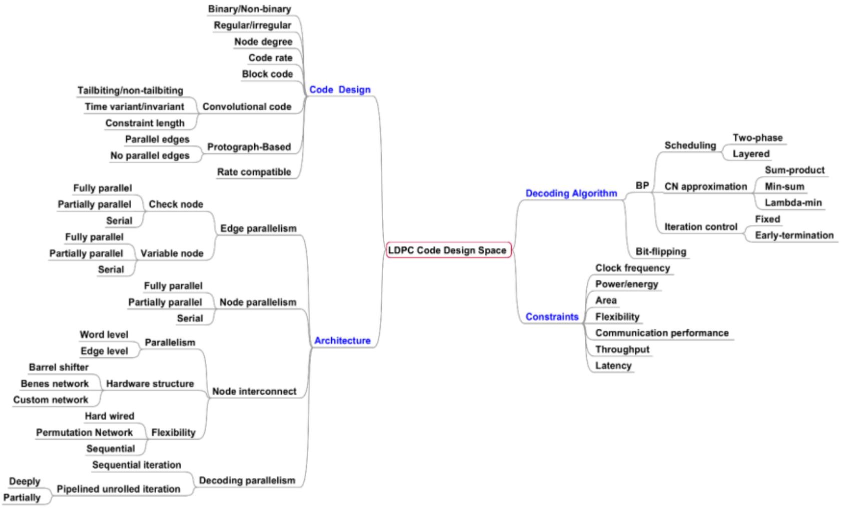
➔Design space exploration

Design Space Turbo Codes



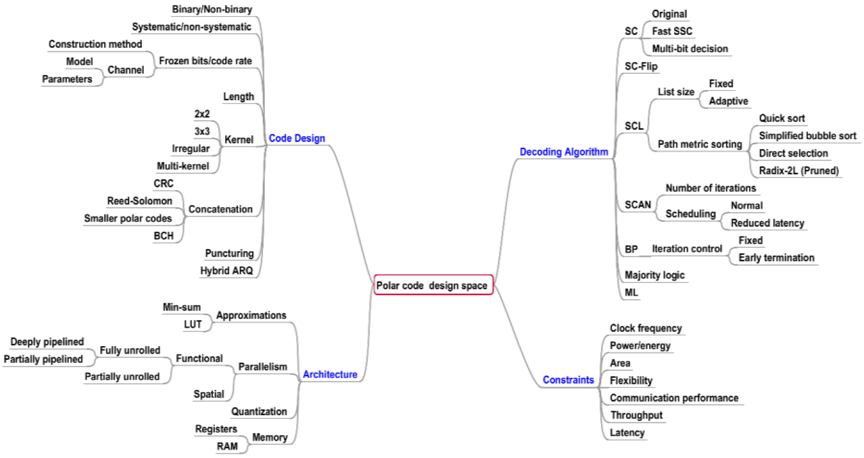
- Turbo code design space
  - Code Design
    - Concatenation
      - Parallel
      - Serial
      - Hybrid
      - Outer code
        - CRC
        - BCH
        - RS
    - Component codes
      - Block-based
      - Convolutional-based
        - Generator polynomials
        - Binary/Non-binary
        - Trellis termination
    - Interleaving
      - Random-based
      - Algebraic
        - QPP
        - ARP
        - DRP
    - Rate adaptation
      - Periodic
      - Incremental
      - Unequal error protection
  - Architecture
    - Serial/parallel
      - Serial
      - Parallel
        - System level
        - Turbo decoder level
          - Shuffled decoding
          - Sequential decoding
            - Unrolled pipelined
            - Serial processing
        - Component decoder level
          - RAM technology
          - Parallel processing
          - Modified scheduling
        - Recursion level
          - Serial
          - Pipelined
          - Unrolled
        - Trellis-stage level
          - Radix
          - Trellis compression
          - Pipelined
    - Interleaver addresses
      - Memory-based
      - Address generator
      - Parallel processing
        - Contention-free
        - RAM permutation
        - Conflict resolution
          - Run time
          - Design time
    - Metrics
      - Metric scaling and quantization
      - Metrics management
        - Recomputation
        - Storing
      - State metric initialization
        - Next Iteration Initialization (NII)
        - ACQuisition
        - Hybrid NII + ACQ
  - Decoding Algorithm
    - Max-Log MAP
      - Sliding window
        - Window size
        - Initialisation
          - NII
          - ACQ
          - NII+ACQ
      - Fully parallel
        - Information exchange management
        - No management
      - With Jacobian logarithm correction
      - With scaling factor
      - Stopping criteria
        - Max number of iterations
        - On-the-fly CRC
        - Extrinsic info-based
    - SOVA
      - Reduced complexity SOVA
      - Trimming SOVA
    - BP
      - Parity-check matrix modification
      - Detection of error-prone positions
    - Bit-flipping
      - Event flipping
      - Improved flip and check
    - ML
    - Stochastic
    - Modified Trellis
      - Trellis definition
      - Equivalency management
    - Saturation
      - Improved symbol forcing
      - Correction impulse
  - Constraints
    - Clock frequency
    - Power/energy
    - Area
    - Flexibility
    - Communication performance
    - Throughput
    - Latency

Design Space LDPC Codes



- LDPC Code Design Space
  - Code Design
    - Binary/Non-binary
    - Regular/irregular
    - Node degree
    - Code rate
    - Block code
    - Convolutional code
      - Tailbiting/non-tailbiting
      - Time variant/invariant
      - Constraint length
    - Protograph-Based
      - Parallel edges
      - No parallel edges
    - Rate compatible
  - Architecture
    - Edge parallelism
      - Check node
        - Fully parallel
        - Partially parallel
        - Serial
      - Variable node
        - Fully parallel
        - Partially parallel
        - Serial
    - Node parallelism
      - Fully parallel
      - Partially parallel
      - Serial
    - Node interconnect
      - Parallelism
        - Word level
        - Edge level
      - Hardware structure
        - Barrel shifter
        - Benes network
        - Custom network
      - Flexibility
        - Hard wired
        - Permutation Network
        - Sequential
    - Decoding parallelism
      - Sequential iteration
      - Pipelined unrolled iteration
        - Deeply
        - Partially
  - Decoding Algorithm
    - BP
      - Scheduling
        - Two-phase
        - Layered
      - CN approximation
        - Sum-product
        - Min-sum
        - Lambda-min
      - Iteration control
        - Fixed
        - Early-termination
    - Bit-flipping
  - Constraints
    - Clock frequency
    - Power/energy
    - Area
    - Flexibility
    - Communication performance
    - Throughput
    - Latency

Design Space Polar Codes



- Polar code design space
  - Code Design
    - Binary/Non-binary
    - Systematic/non-systematic
    - Frozen bits/code rate
      - Construction method
      - Channel
        - Model
        - Parameters
    - Length
    - Kernel
      - 2x2
      - 3x3
      - Irregular
      - Multi-kernel
    - Concatenation
      - CRC
      - Reed-Solomon
      - Smaller polar codes
      - BCH
    - Puncturing
    - Hybrid ARQ
  - Architecture
    - Approximations
      - Min-sum
      - LUT
    - Parallelism
      - Functional
        - Fully unrolled
          - Deeply pipelined
          - Partially pipelined
        - Partially unrolled
      - Spatial
    - Quantization
    - Memory
      - Registers
      - RAM
  - Decoding Algorithm
    - SC
      - Original
      - Fast SSC
      - Multi-bit decision
    - SC-Flip
    - SCL
      - List size
        - Fixed
        - Adaptive
      - Path metric sorting
        - Quick sort
        - Simplified bubble sort
        - Direct selection
        - Radix-2L (Pruned)
    - SCAN
      - Number of iterations
      - Scheduling
        - Normal
        - Reduced latency
    - BP
      - Iteration control
        - Fixed
        - Early termination
    - Majority logic
    - ML
  - Constraints
    - Clock frequency
    - Power/energy
    - Area
    - Flexibility
    - Communication performance
    - Throughput
    - Latency

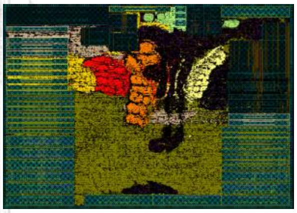
Compute vs Data Transfer/Storage

E.g. Belief propagation

- Inherent parallel (check/variable node processing)
- Data transfer dominated

Data transfers

- Highly parallel architectures -> wires
  - routing congestion: area and power



Design Step	Area mm²	Clock Frequency	Power
Synthesis	2,9	322 MHz	~ 600 mW
P & R	4,6	275 MHz	1110 mW

Large difference: area 58%, power 83% increase, frequency 17% decrease

Compute vs Data Transfer/Storage

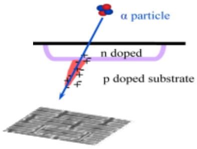
High throughput, (partially) parallel architectures: SRAM memories

- Largely contribute to power

Memory	
Cache	(64bit)
8KB	10pJ
32KB	20pJ
1MB	100pJ

Integer	
Add	
8 bit	0.03pJ
32 bit	0.1pJ
Mult	
8 bit	0.2pJ
32 bit	3.1pJ

- Generate access conflicts
  - E.g. TC interleaver, double diagonal Q matrix LDPC…
- Reliability issue due to high energy particles
  - ECC protection becomes mandatory

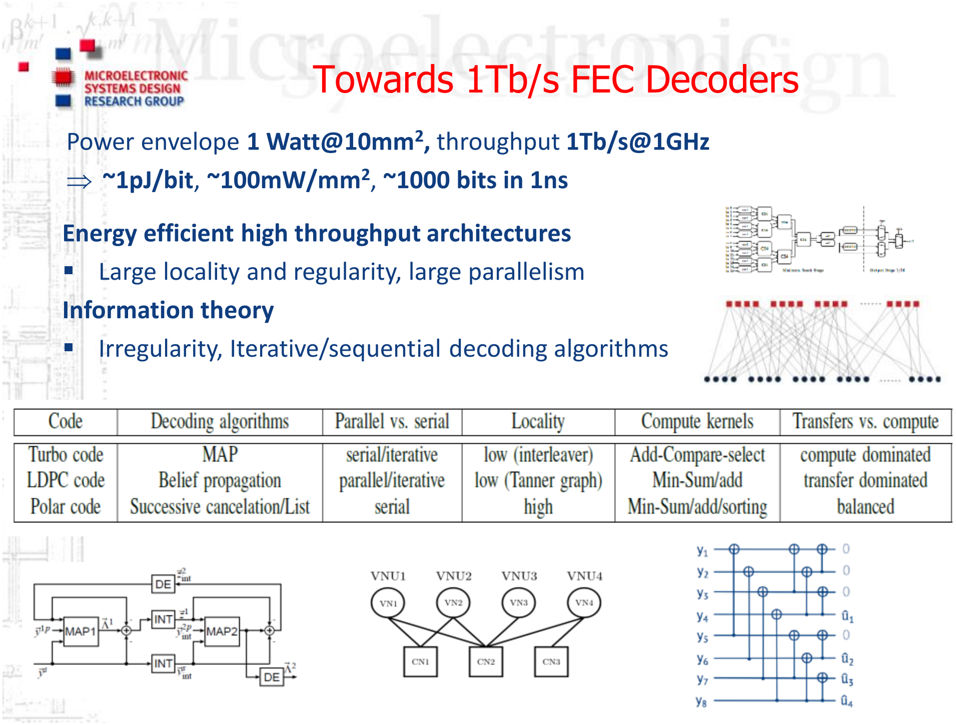


Complexity of data transfer (routing) / storage **~ N x P x q**

- N blocklength, P parallel decoded codewords, q average LLR precision

=> Efficient LLR quantization q is a major optimization step

E.g. information bottleneck, finite alphabets…



MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

Towards 1Tb/s FEC Decoders

Power envelope **1 Watt@10mm^2,** throughput **1Tb/s@1GHz**

⇒ **~1pJ/bit**, **~100mW/mm^2**, **~1000 bits in 1ns**

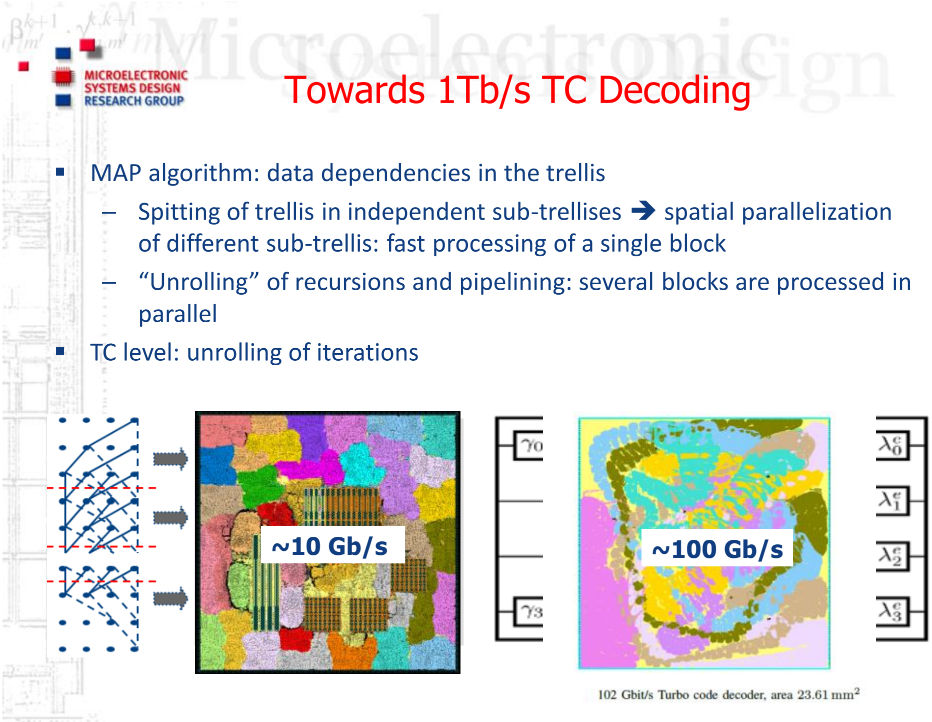
Energy efficient high throughput architectures

- Large locality and regularity, large parallelism

Information theory

- Irregularity, Iterative/sequential decoding algorithms

Code	Decoding algorithms	Parallel vs. serial	Locality	Compute kernels	Transfers vs. compute
Turbo code	MAP	serial/iterative	low (interleaver)	Add-Compare-select	compute dominated
LDPC code	Belief propagation	parallel/iterative	low (Tanner graph)	Min-Sum/add	transfer dominated
Polar code	Successive cancelation/List	serial	high	Min-Sum/add/sorting	balanced



MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

Towards 1Tb/s TC Decoding

- MAP algorithm: data dependencies in the trellis
  - Spitting of trellis in independent sub-trellises ➔ spatial parallelization of different sub-trellis: fast processing of a single block
  - "Unrolling" of recursions and pipelining: several blocks are processed in parallel
- TC level: unrolling of iterations

102 Gbit/s Turbo code decoder, area 23.61 mm^2

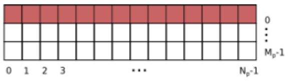
Towards 1Tb/s LDPC Decoding

LDPC Block code length N, parity check matrix H, I iterations
#edges: 1_entries(H), *#proc_edges(H)* edges processed in 1 clock cycle

Parallelism P

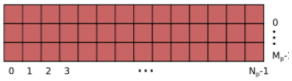
$$T_{BC}(H,A) = \frac{\#proc\_edges(A)}{\#edges(H)} * N * \frac{1}{i} * f \quad \text{[bits/sec]}$$

P<1 (i=I)



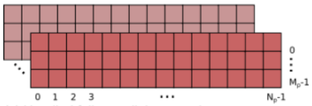
(a) Partially parallel processing

P=1 (i=I)



(b) Fully parallel processing

P>1 (i=1)



(c) Unrolled fully parallel processing

E.g. IEEE 802.11ad, N=672, #edges(H)=1890, f=400MHz (28nm FDSOI), 9 iter.

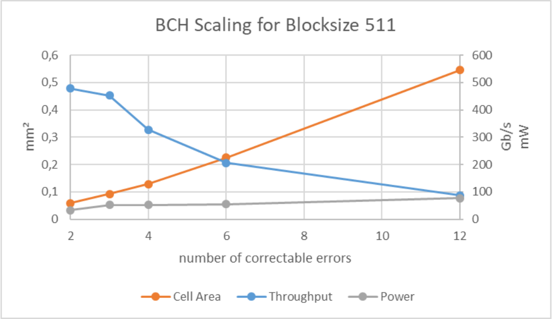
- ~10 Gb/s P<1 e.g. row wise partially parallel architecture (10 Gbit/s)
- >10 Gb/s P=1 i.e. fully parallel on Tanner graph (29 Gbit/s)
- >100 Gb/s P>1 i.e. several H matrices are processed in parallel, unrolled fully parallel architecture (268 Gbit/s)

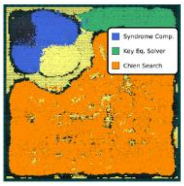
268 Gbit/s LDPC code decoder, area 2.8 mm²

BCH Decoder Scaling

Extended Euclidian Algorithm, piplined architecture
28nm FDSOI, WC PVT, blocksize 511, correctable errors: 2..12

Place & Route	2	3	4	6	12
Frequency (MHz)	935	885	641	403	170
Throughput (Gbps)	477,6	452,2	327,6	206	86,8
Total Cell Area (um2)	59198	93014	128475	224906	545406
Area Efficiency (Gbps/mm2)	5247	3397	1793	670	119
Power Total (mW)	33,42	52,7	52,24	54,44	76,57
Energy Efficiency (pJ/bit)	0,07	0,12	0,16	0,26	0,88

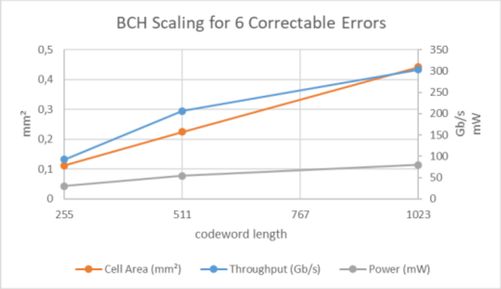




BCH Decoder Scaling

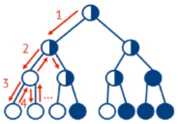
28nm FDSOI, WC PVT, 6 correctable errors, blocksize: 255…1023

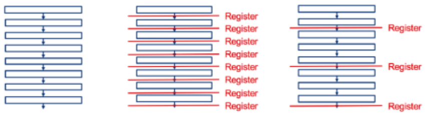
Place & Route	255	511	1023
Frequency (MHz)	364	403	297
Throughput (Gbps)	92,7	206	303,6
Total Cell Area (um2)	111902	224906	442433
Area Efficiency (Gbps/mm2)	617	670	481
Power Total (mW)	30,15	54,44	79,82
Energy Efficiency (pJ/bit)	0,33	0,26	0,26



Towards 1Tb/s Polar Decoding

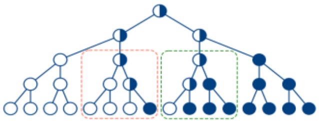
- Decoding algorithms SC, SCL: “unrolling” of tree traversal on polar factor tree



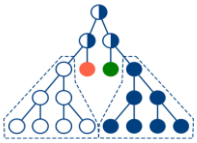


2*(2N-2)+1 stages

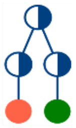
- Reduction of tree size by different optimizations e.g.
  – Replace repetition codes and parity check code by one single nodes
  – Merge rate-0 codes and rate-1 nodes into parent nodes



Original Tree



Replaced subtrees



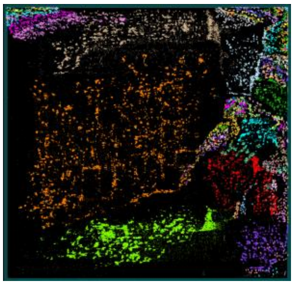
Optimized tree

Towards 1Tb/s SC Polar Decoding

1024/512 Code, fast SC decoding algorithms

- Worst case PVT timing 28nm technology, optimized factor tree,
- Logic stages 385, retimed pipeline stages 105

Place&Route	Register	Latches
Area [mm^2]	3.14	2.79
- Combinat.	0.96	0.91
- Buf/Inv	0.65	0.27
- Noncomb	1.55	1.12
Area Eff. [Gbps/mm^2]	205	231
Utilization	78%	72%
Frequency [MHz]	621	629
Throughput [Gbps]	636	644
Power [W]	5.7	2.7
- Clock	47%	19%
- Registers	24%	13%
- Combinat.	29%	68%
Energy Eff. [pJ/bit]	8.8	4.2



Each colour represents a stage (105)
black color is memory

Summary

- Applications require ever higher throughput, lower latency, better communication performance, higher energy efficiency and low power
- Microelectronic progress can not keep pace with these requirements
- Throughput towards 1 Tb/s are feasible for TC, LDPC, PC but
  - Limited to smaller block sizes, low iterations (TC, LDPC) ➔ comm. performance
  - Flexibility challenge
  - Heavy pipelining increases latency, power in clock tree is a major challenge
  - Power density one of the biggest challenges
- High communication performance under architectural constraints for very high throughput is challenging

Thank you for attention!

For more information please visit

http://ems.eit.uni-kl.de