IBM Research

Multi-Level Design of Energy-Efficient Adaptive Networks for Large Computing Systems

Juan-Antonio Carballo
IBM Research

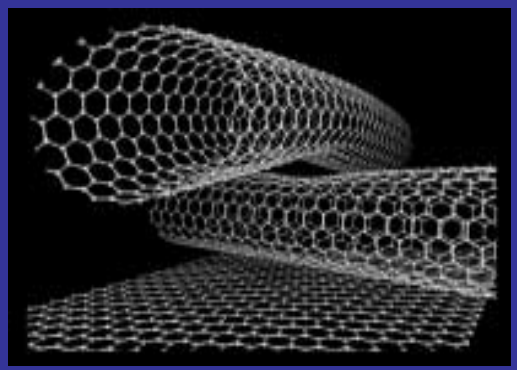
jantonio@us.ibm.com

April 2004

IBM Research Worldwide

Research's Strategic Thrusts → Vital to IBM's Future

Exploratory Science

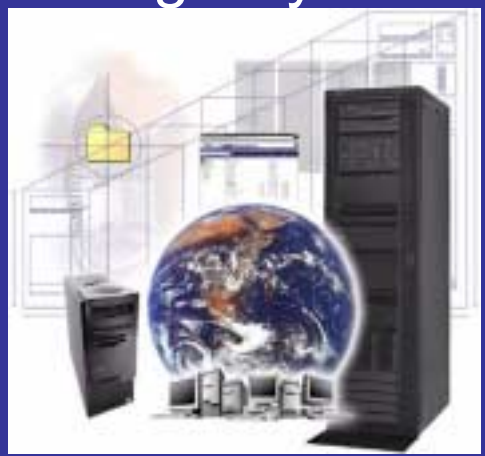


Servers & Embedded Systems

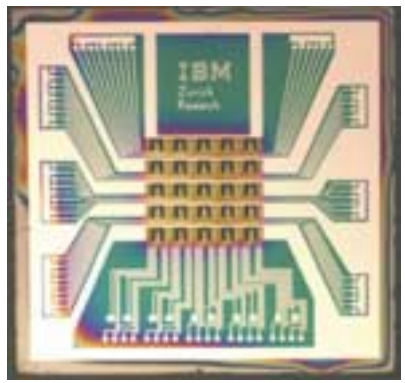
Personal Systems

Services & Software

Storage Systems



Technology



Overview

- **Future large computing systems: the PERCS project**
  - Vision → adaptability, productivity, performance
  - Scope → application focus, integration
- **The importance of networking and communications**
  - Power as the new performance
- **Energy-efficient, productive interconnect design**
  - Multi-level communications design optimization
  - Adaptive communications architectures

PERCS → Design Constraints

- **Legacy investments**
- **Looming technology crisis**
- **HPC customer diversity**
- **Business model**

  Must do well both on commercial and scientific workloads
- **Cost issues**

  Threat of commoditization
- **Productivity as a main theme**

IBM's Vision

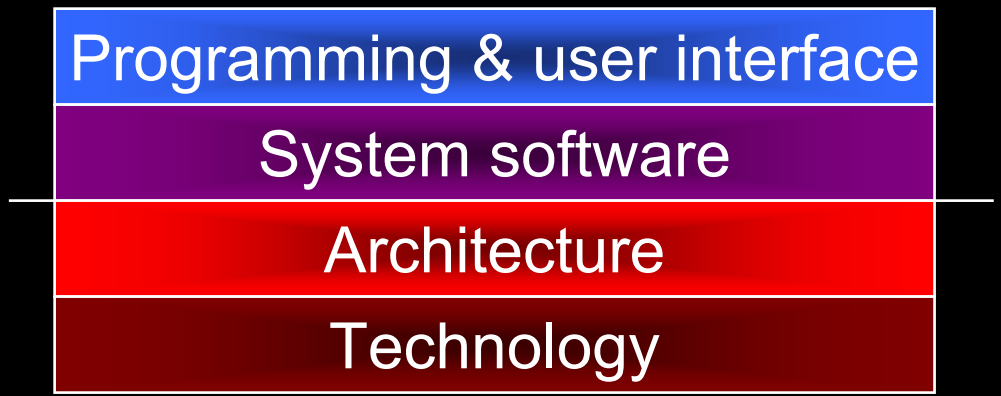
A dynamic system that adapts to application needs

The strategy

- **Aggressive productivity targets**
- **Commercial viability**
- **Link into product cycle toward end of phase 2**

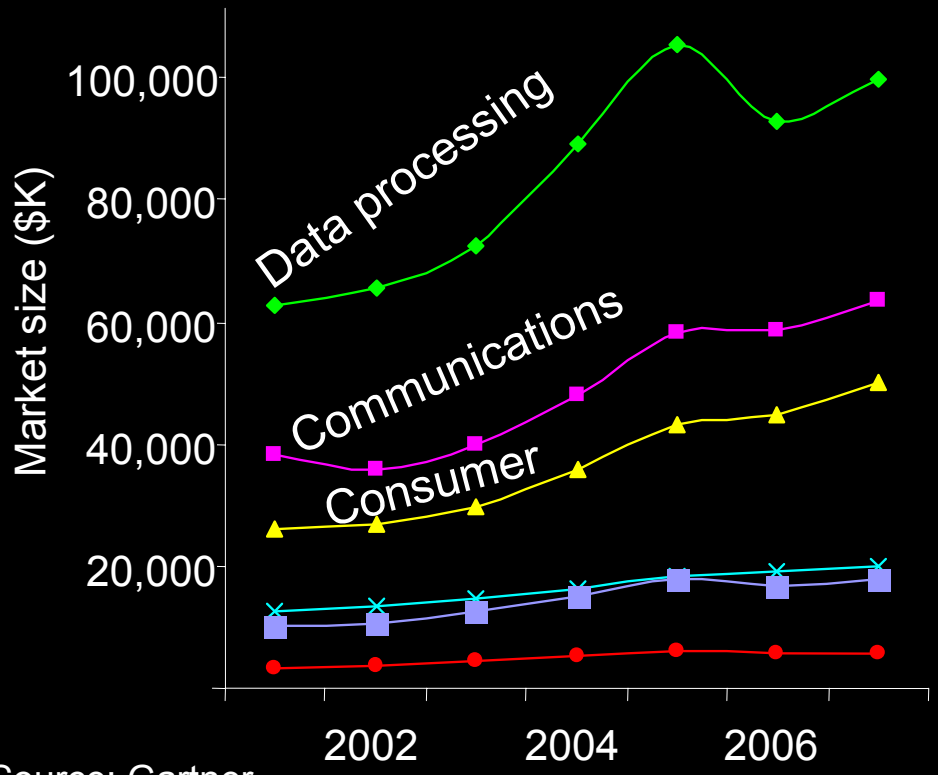
PERCS → Scope

- **Application focus**
  - Commercial
  - Security
  - Bioinformatics
  - Data streaming
  - New 2010-apps ??
- **Integrated solution**
  - HPC

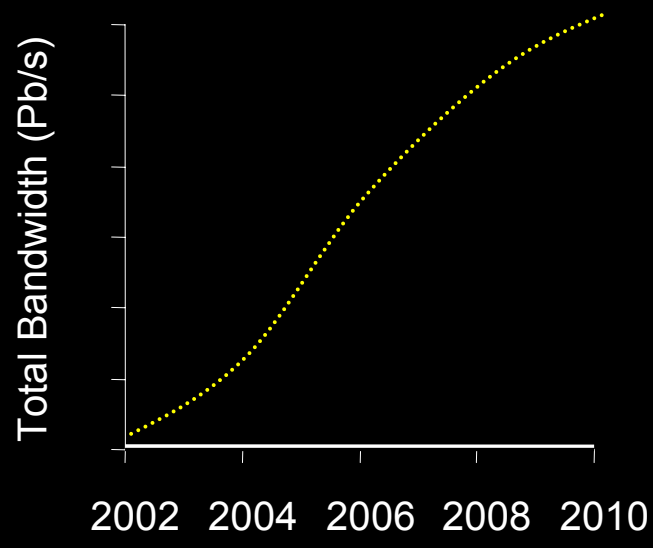


Importance of Communications Hardware

- **Important as a chip market**
  One of industry's key segments
- **Key to server / large computers**
  High-BW: distinguishing feature



Source: Gartner



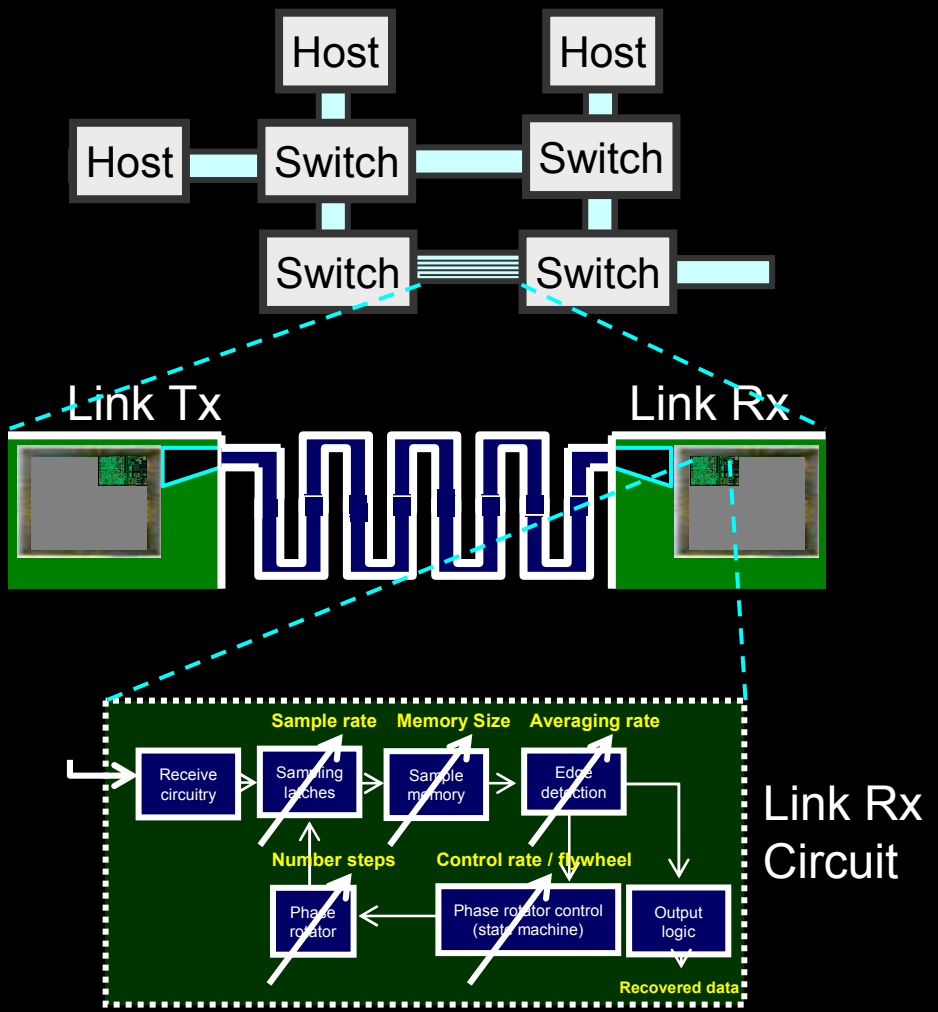
Source: IBM large computing projects

A Multi-Level Power Management Problem

Network design — **W**

Link design — **mW**

Circuit design — **µW**



Designing High-Bandwidth Links is Difficult

- **Challenges**
  - High-speed, low-power, low-BER, many customers
  - Uncertainties in package, channel, chip manufacturing
  - Mixed-signal system sensitive to these uncertainties

Communications in HPC → Key Questions

- **Goal: to achieve**
  - High productivity (use, design)
  - Commercial viability
  - Competitive performance at acceptable power

Communications in HPC → Key Focus Areas

1. Adaptability to application requirements

Workloads, protocols, channels, speeds

→ viability (applications), productivity (reuse)

2. Adaptability to environment variations

Manufacturing quality, post-manufacturing conditions

→ viability (yield), productivity (design)

3. Hardware design productivity

Single design, single design methodology

→ productivity (design)

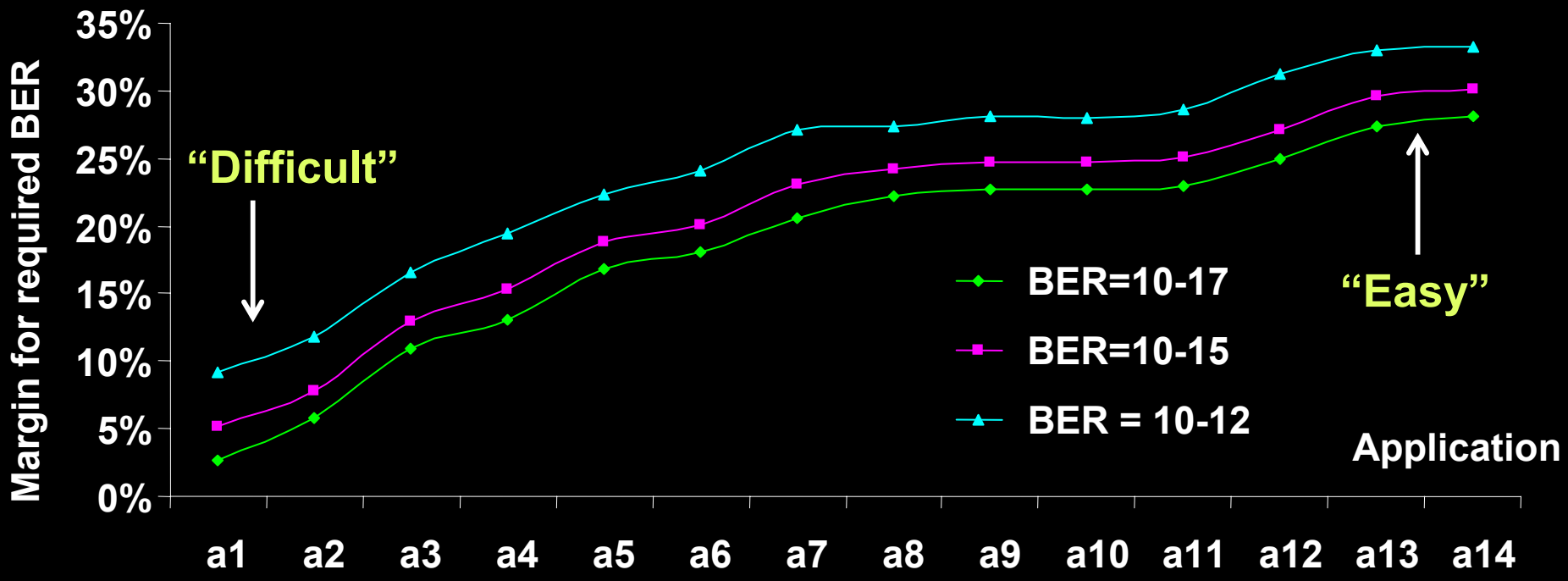
Multi-Level Communications Design

Level	Design strategy	Adaptability	Productivity
Link design	Self-adaptive links	Application, environment 50% better power	Single design
	On-core problem determination	Post-manufacturing environment	Debugging
Circuit design	1/2-custom islands	Supply variations 25% better power	Design convergence
	Custom digital Custom analog	Manufacturing variations 10% better power	Design convergence

1. Adaptive Communications Links

- **Application requirements, environment impact performance**

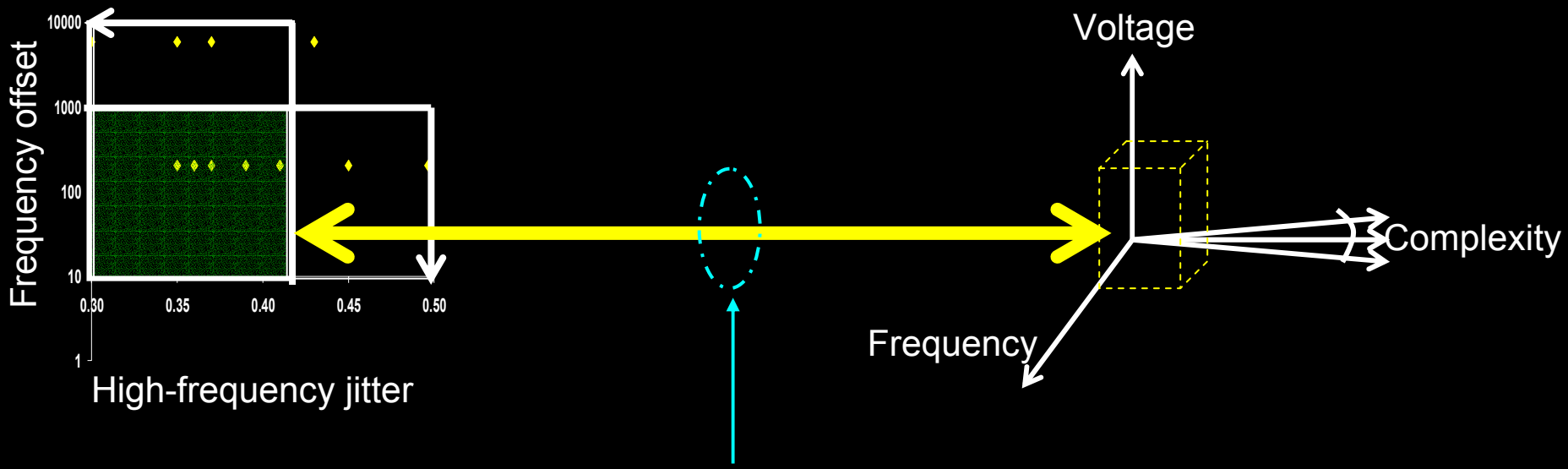
  Workload, frequency, quality of channel, package, chip layout



1. "Difficult" Versus "Easy" Requirements

	Difficult	Easy
Workload	Fast-varying frequency pattern	Uniform-frequency pattern
Frequency	High	Low
Channel	Long across boards	Short chip-to-chip
Package	Low-cost low-yield	Ceramic, high quality
Chip layout	Longer wires	Shorter wires

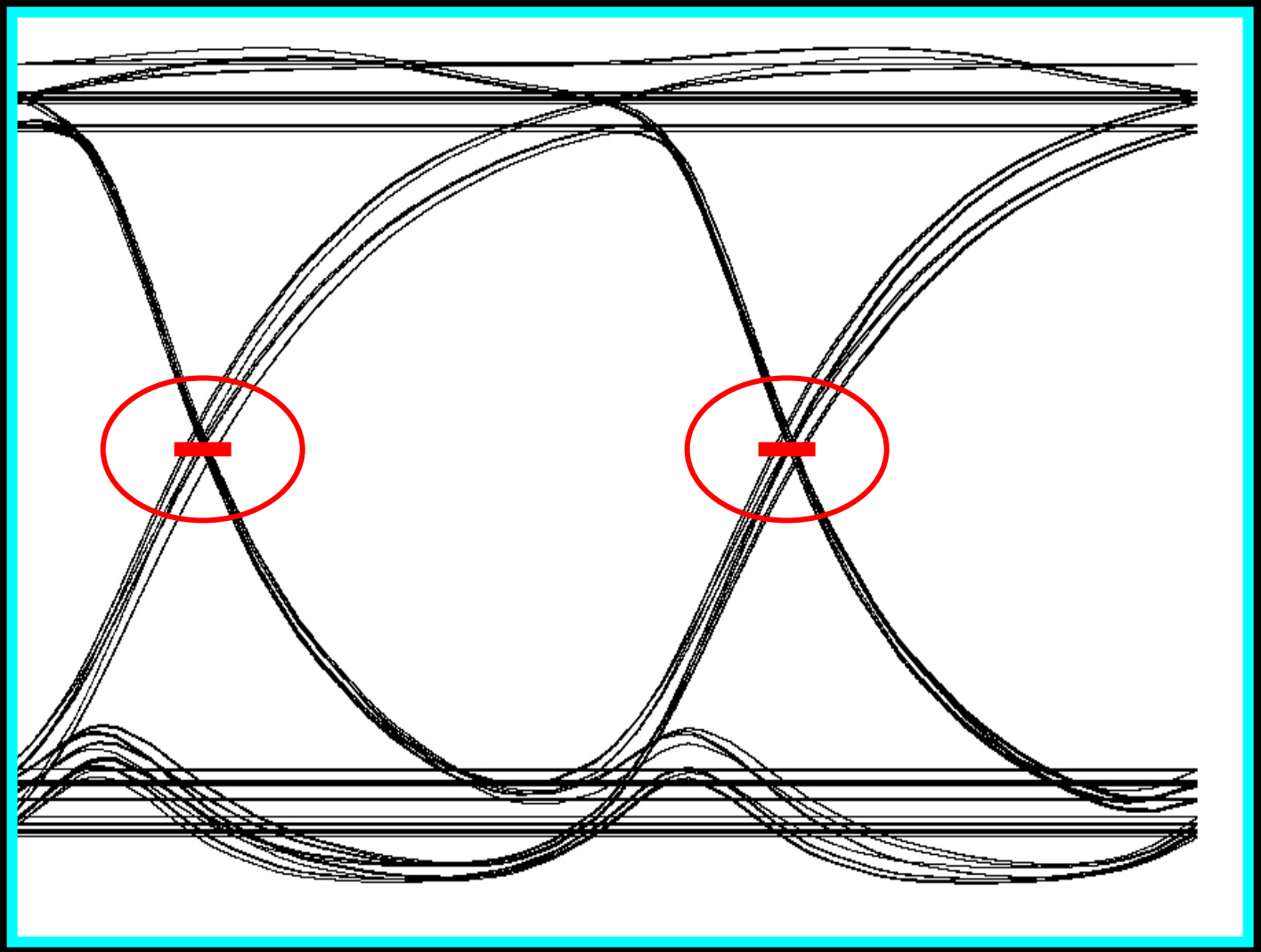
1. Requirement-Based Design of Adaptive Links



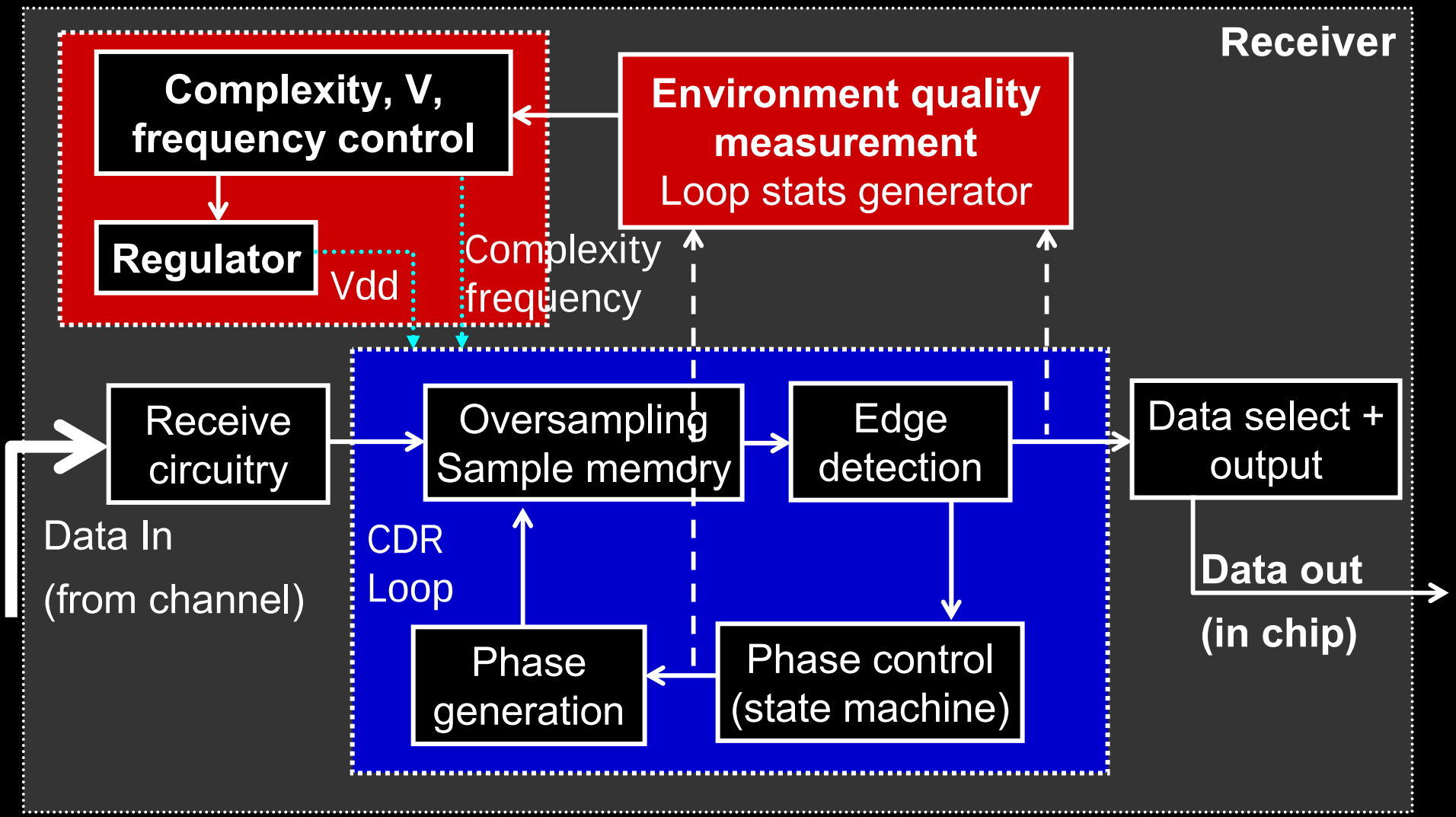
Requirements space ..... **Mapping** ..... **Link Design space**

Requirement measurement blocks	.....	Power modes	.....	Configurable link CDR blocks

1. Jitter (Eye Closure/Movement) Determines Difficulty



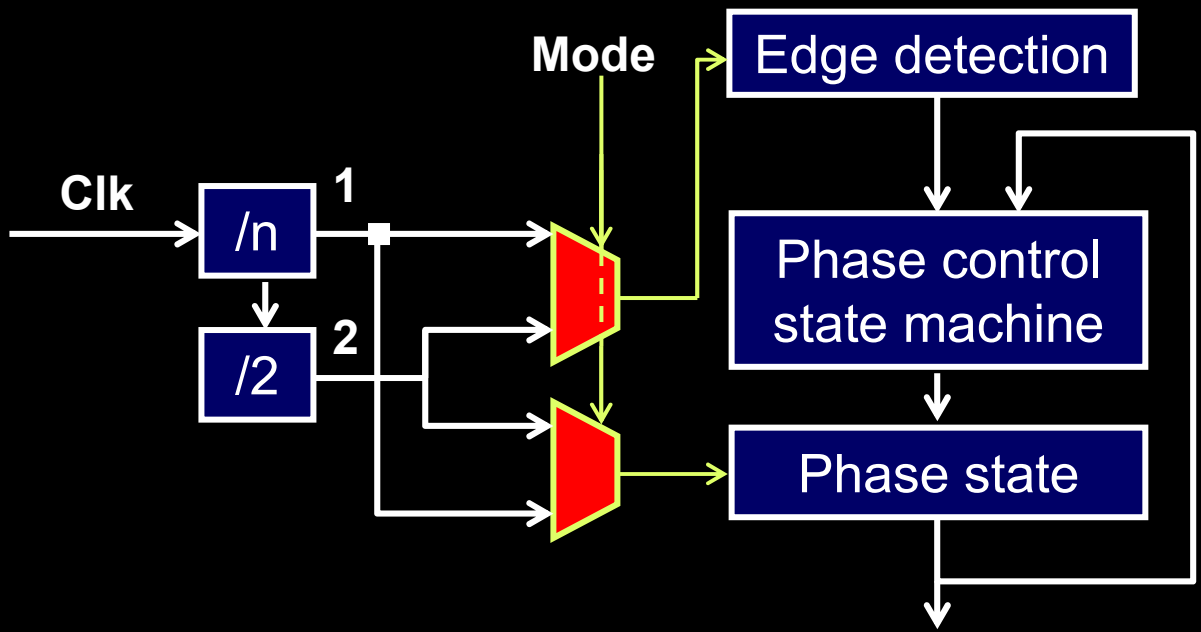
1. Adaptive Link Receiver



1. Example: Adaptive Loop Latency

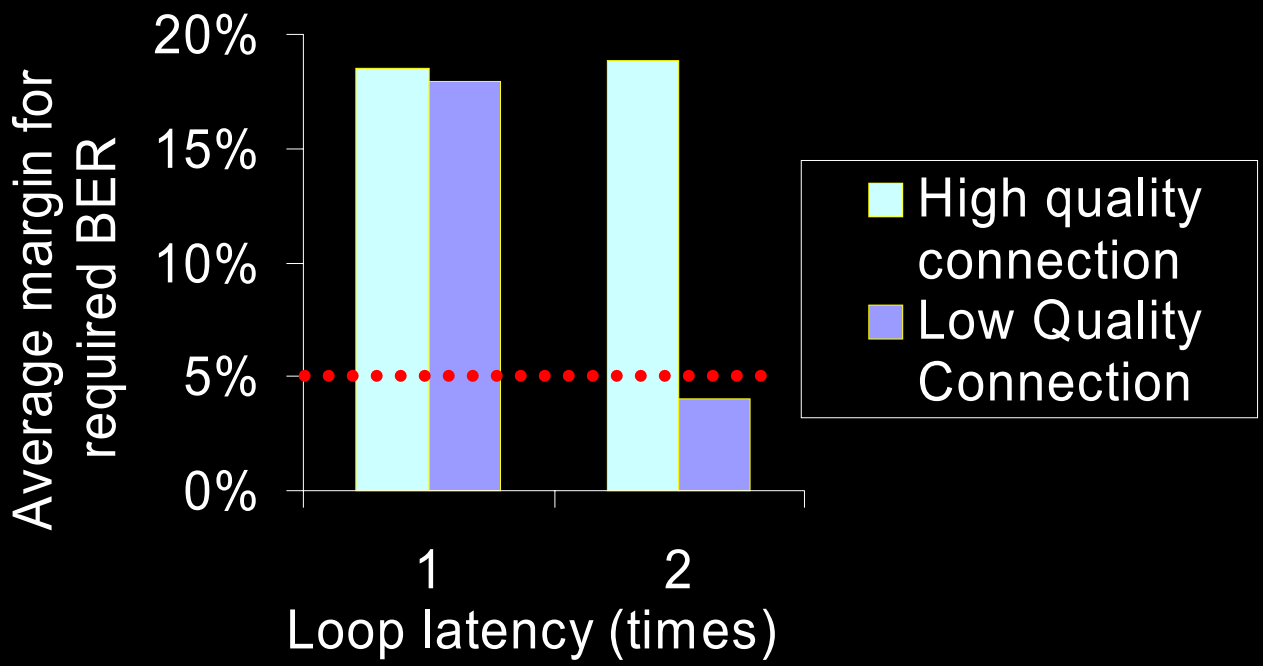
- **Low design overhead**

  Simple clock gating + control logic



1. Adaptability Increases Robustness, Lowers Power

- **Difficult connection → full filter clock speed**
  Low jitter margin → Rx requires fast loop latency
- **High-quality connection → ½ filter clock speed**
  Higher jitter margin → Rx can use low-loop-latency



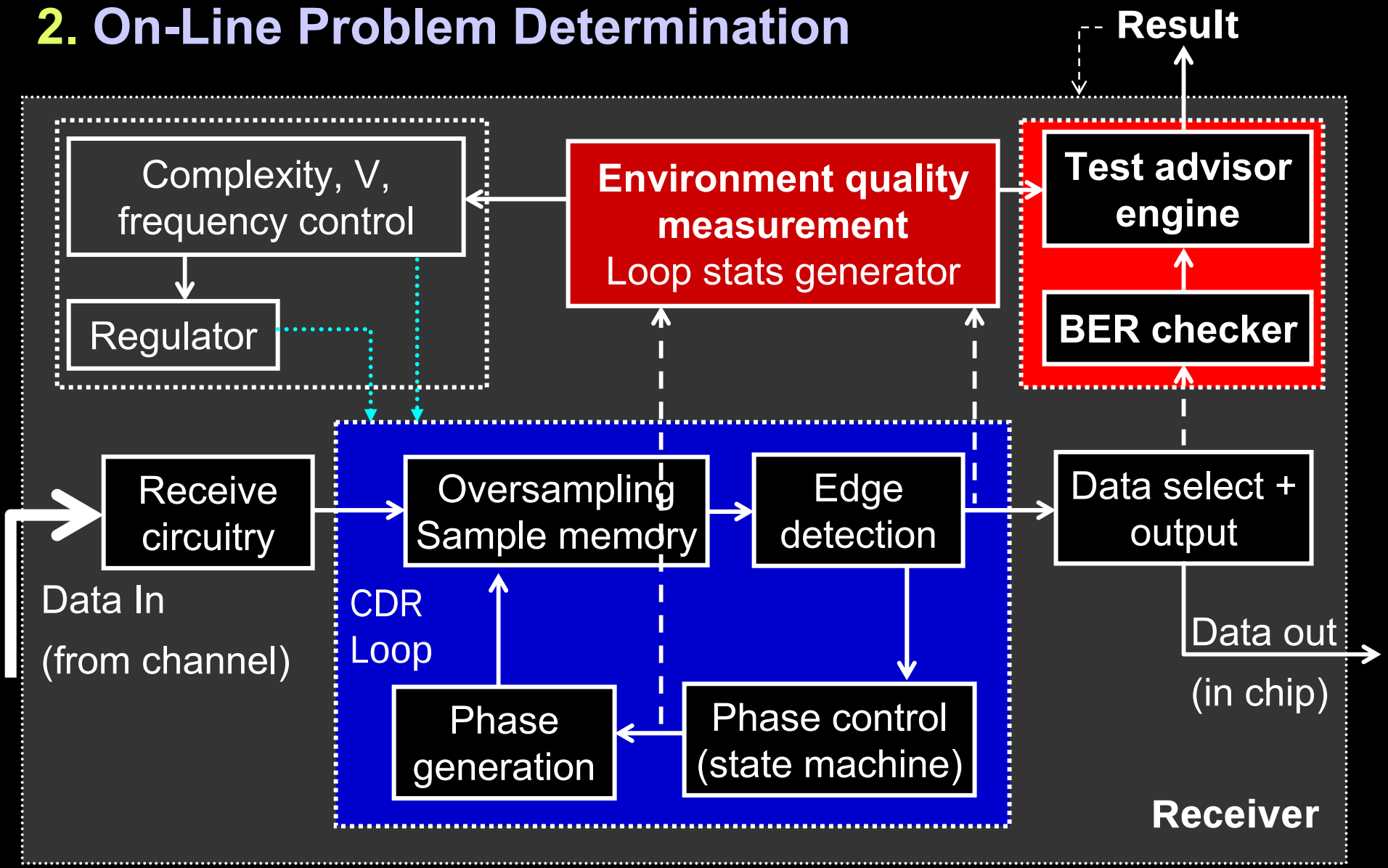
1. Advantages of Adaptive Links

- **Adaptability**
  - To application requirements → automatically minimize power
  - To environment quality → compensate for variations
- **Productivity**
  - Single design, many applications
  - Reduced need for worst-case design
- **Low overhead**
  - Less than 5% area penalty

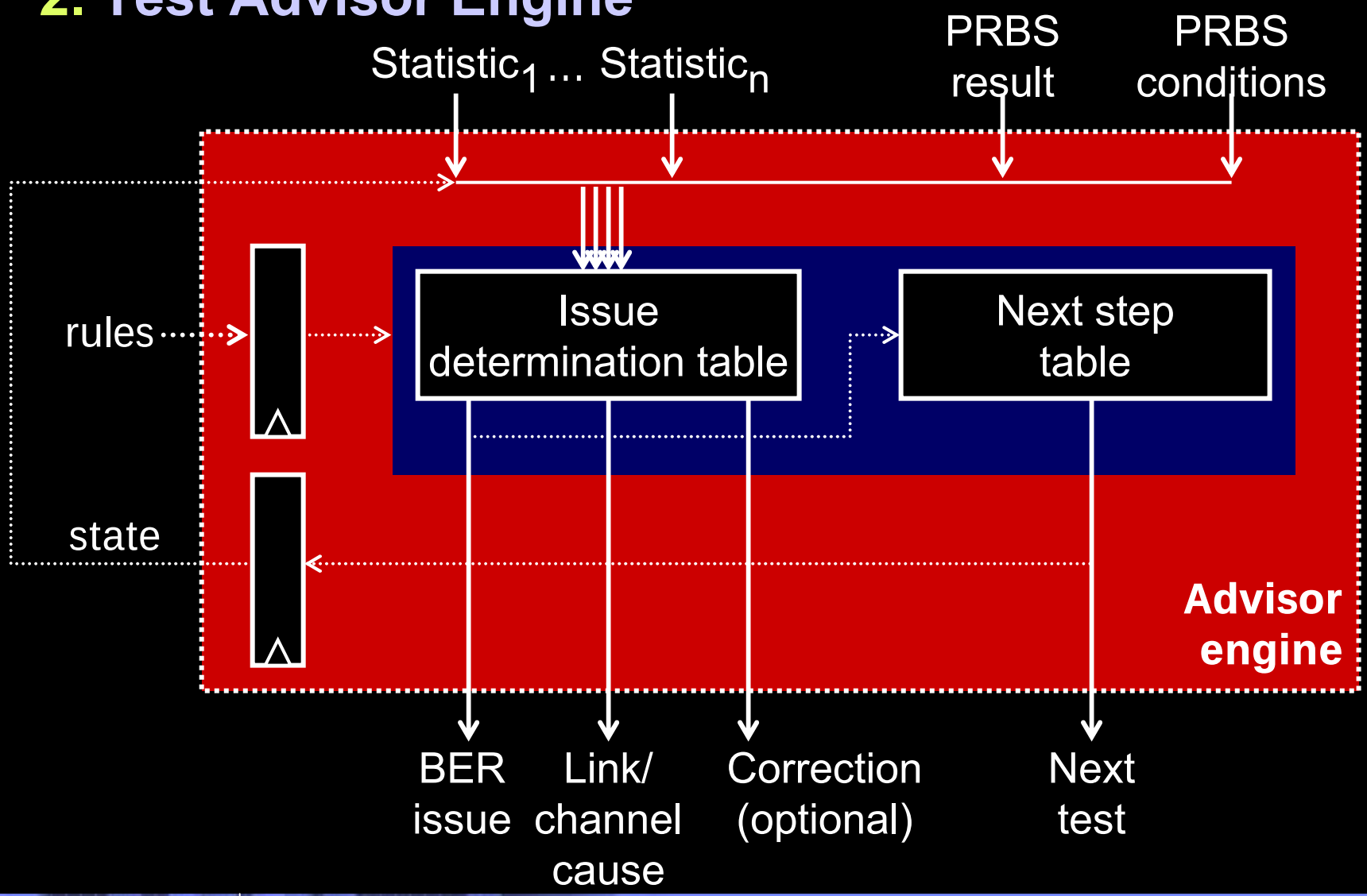
2. On-Line Problem Determination

- **Problem → Once an "adaptable" device is in the field**
  Unexpected design or manufacturing issues may come up
  Must understand environment to effectively configure design
- **Observation → Approach must be**
  capable of determining issue origin (channel or link) and cause
  easy of use, fast in terms of test and/or correction time
- **Solution → *Dual-input* on-line problem determination**
  Combines pattern-based test (1) with analysis of internal signals (2)
  (1) helps understand system performance
  (2) helps understand link behavior

2. On-Line Problem Determination



2. Test Advisor Engine



2. Test Advisor Engine (Example)

Pattern test results		Internal link stats		Outputs of advisor engine		
Pattern	Max. BER	HF Jitter (%UI)	Fr.offset (ppm)	Issue	Cause (s)	Next step
PRBS 7bit	10^{-12}	<50	<200	None	N/A	PRBS 31b
PRBS 31bit	10^{-12}	<50	<200	None	N/A	None
PRBS 31bit	10^{-10}	>75	<200	Channel	H-Freq. jitter (H)	None
PRBS 31bit	10^{-10}	<50	>5000	Application	Freq. offset (H)	None
PRBS 31bit	10^{-10}	<50	<200	Link	Jitter tolerance(L)	JTPAT
JTPAT	10^{-8}	<50	<200	Channel	Group delay (H)	None
JTPAT	10^{-12}	<50	<200	Link	Jitter tolerance(L)	None

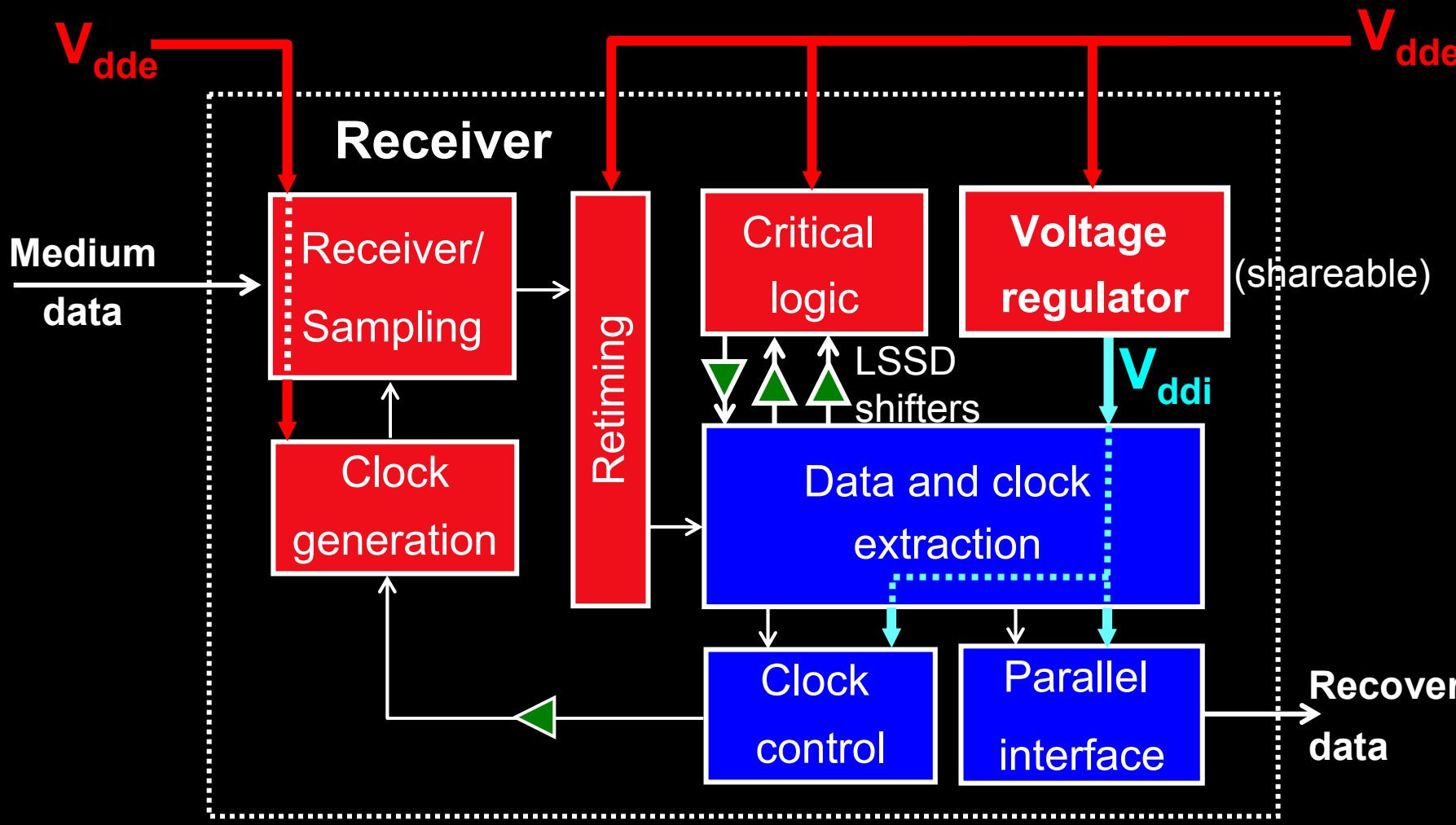
2. Advantages of On-Core Problem Determination

- **Adaptability**
  - To post-manufacturing environment
  - Helps understand what part of link to re-configure and how
- **Productivity**
  - Fast link debugging
  - Low debugging infrastructure cost
- **Low overhead**
  - Less than 2% area penalty

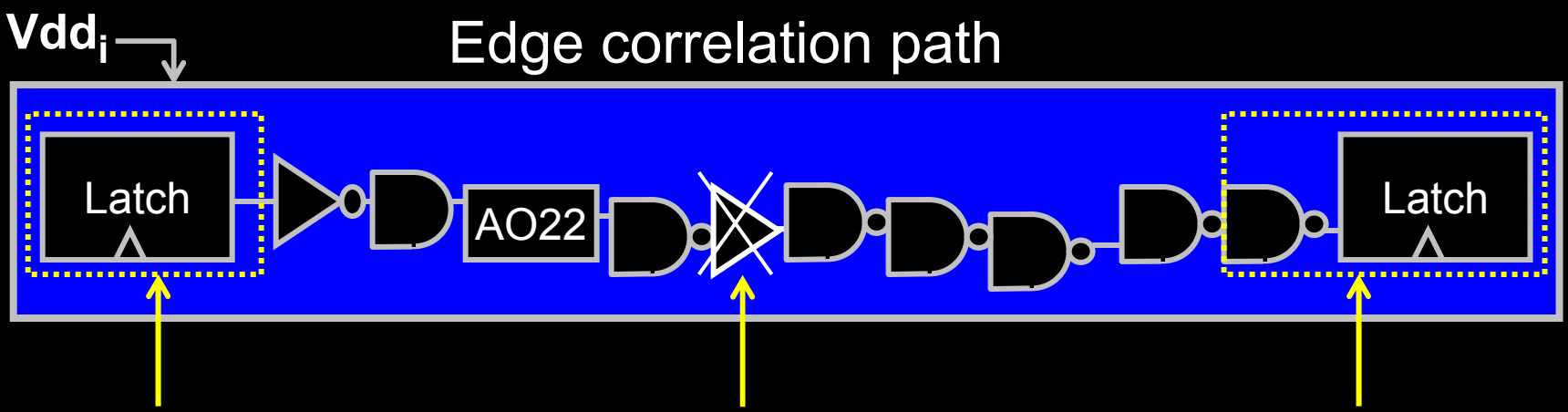
3. Core Design and Integration

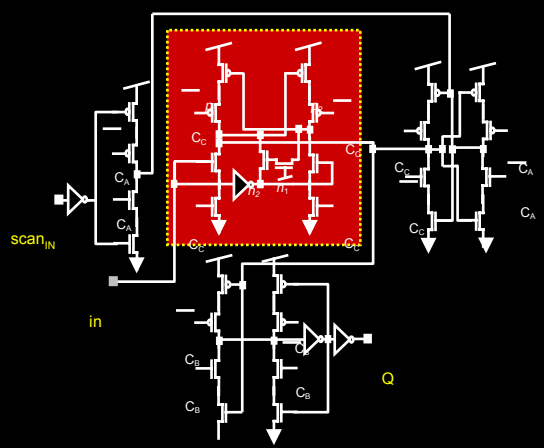
- **Problem: hundreds of high-performance cores**
  - High performance
  - Low power
  - Ease of integration
- **Solution = voltage islands + selective custom design**
  - Performance: custom techniques + multiple V_{th}/V_{dd}
  - Power: low regulated supply
  - Integration: embedded regulation, cell packaging
- **Application: realistically complex links (3000+ gates)**
  - Performance: no impact
  - Power: 25% savings
  - Integration: ASIC methodology, unmodified interfaces

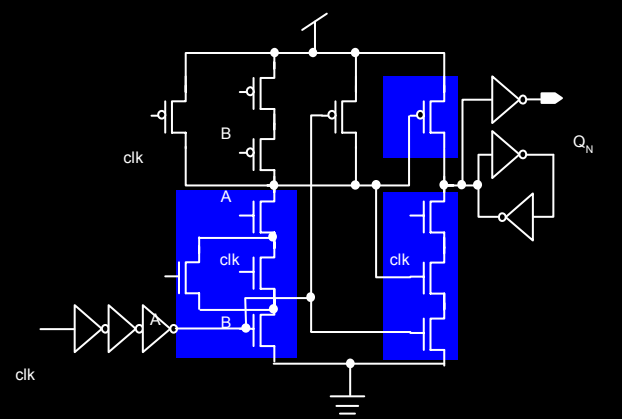
3. Semi-Custom Voltage-Islands



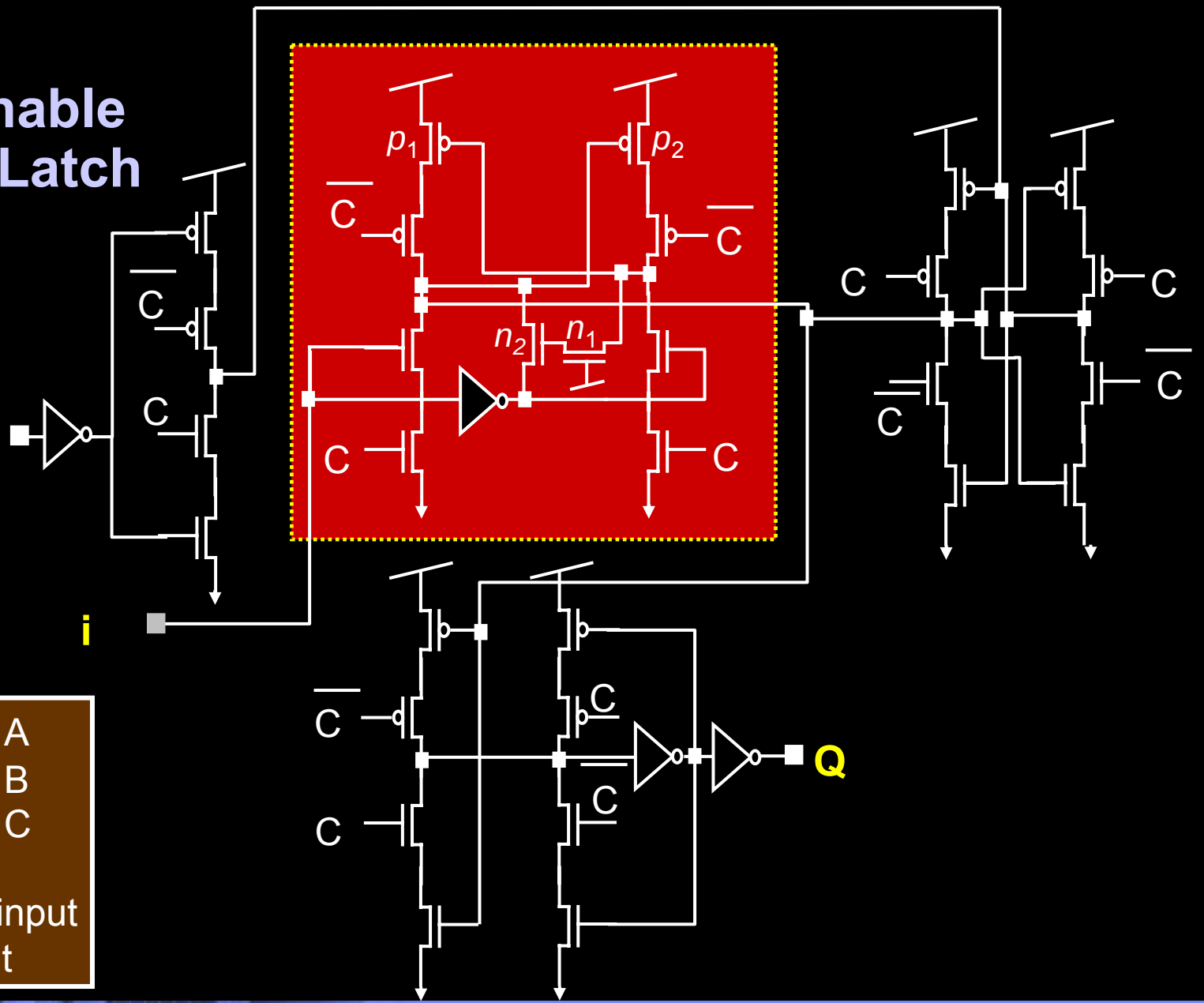
3. Selective Custom Design Increases Flexibility







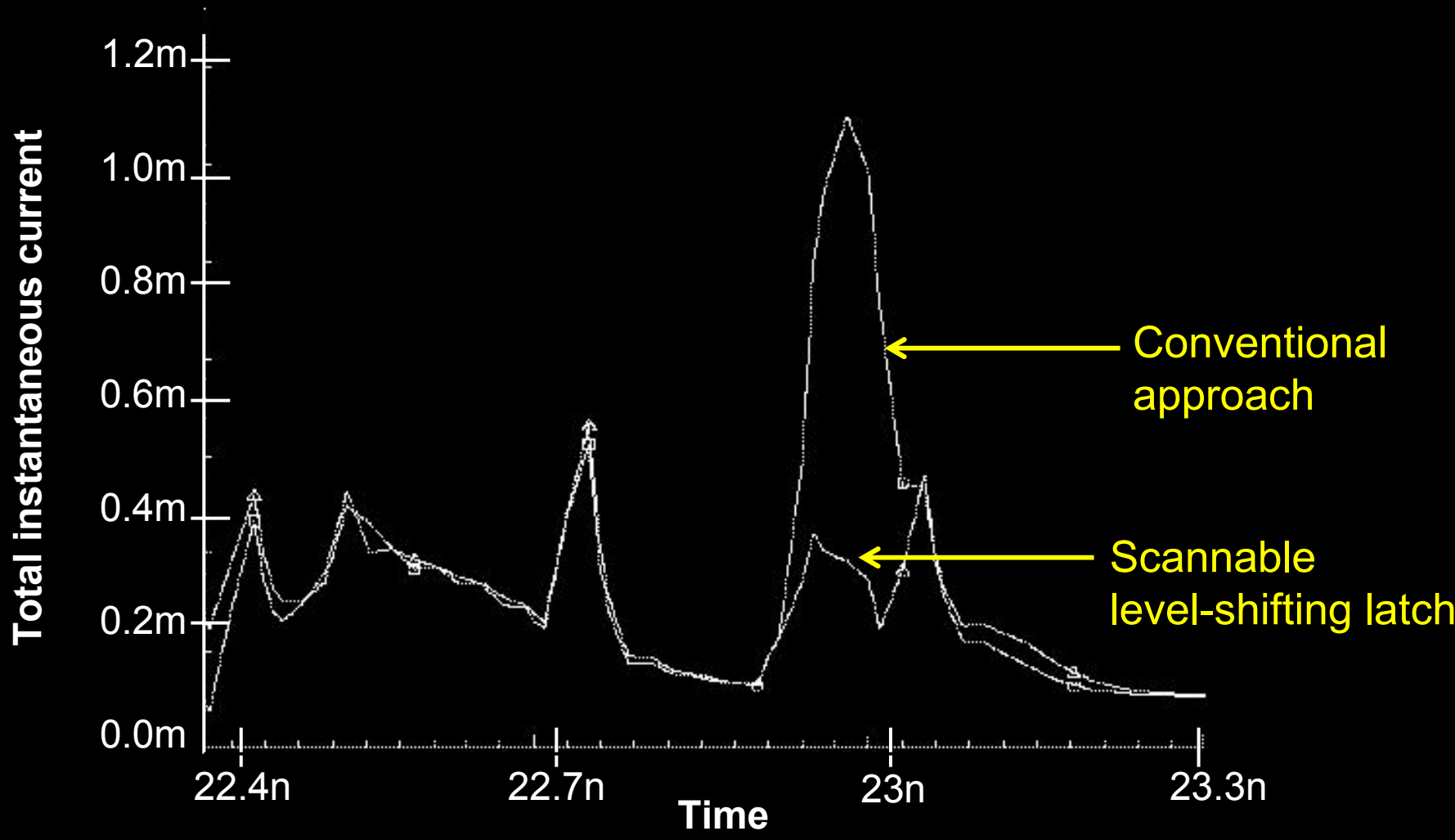
3. Scannable Shifter+Latch



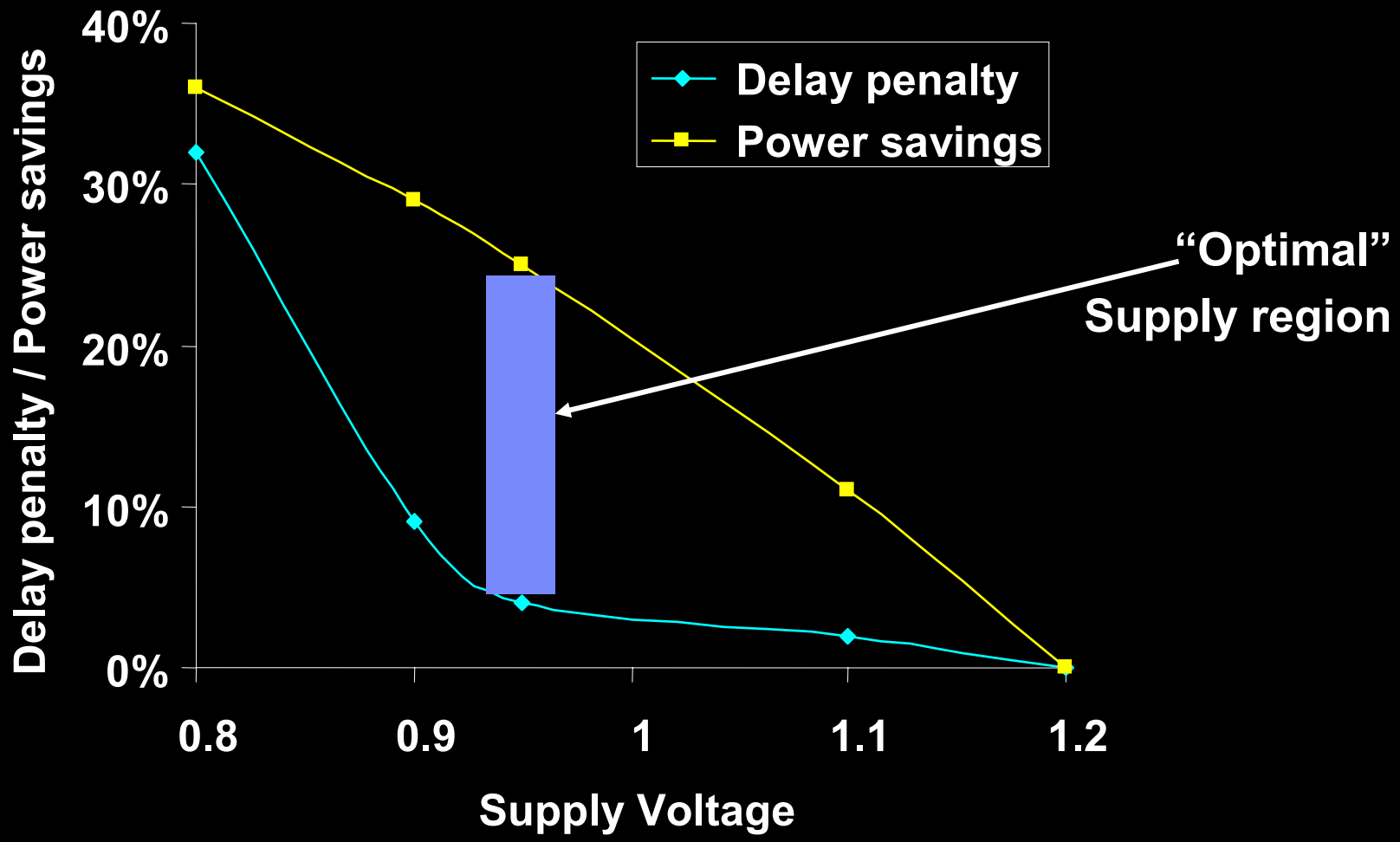
LEGEND

C_A	Clock A
C_B	Clock B
C_C	Clock C
in	Input
$scan_{in}$	Scan input
Q	Output

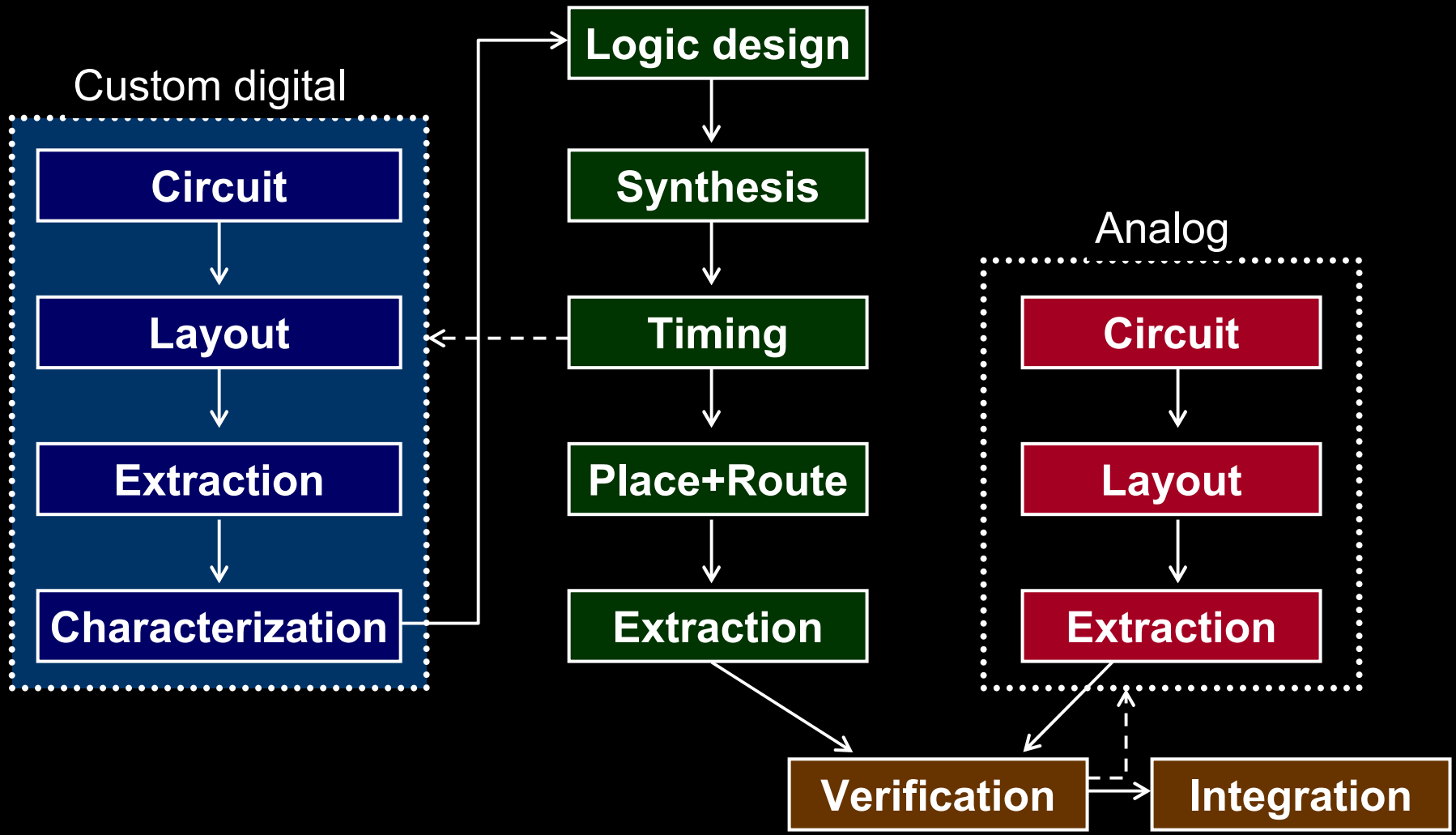
3. Integration/Customization Saves Power



3. Using Critical Paths to Choose Supply Voltage

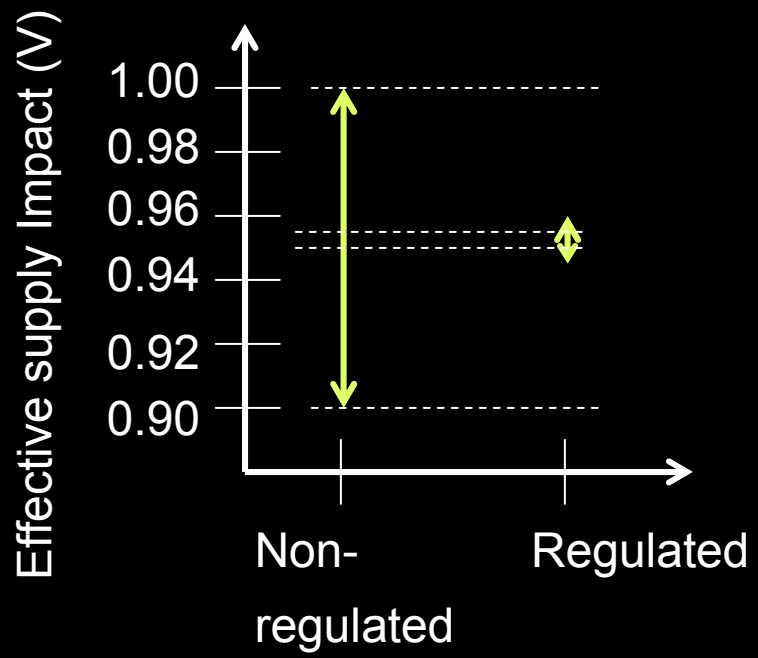


3. Flexible Design Methodology

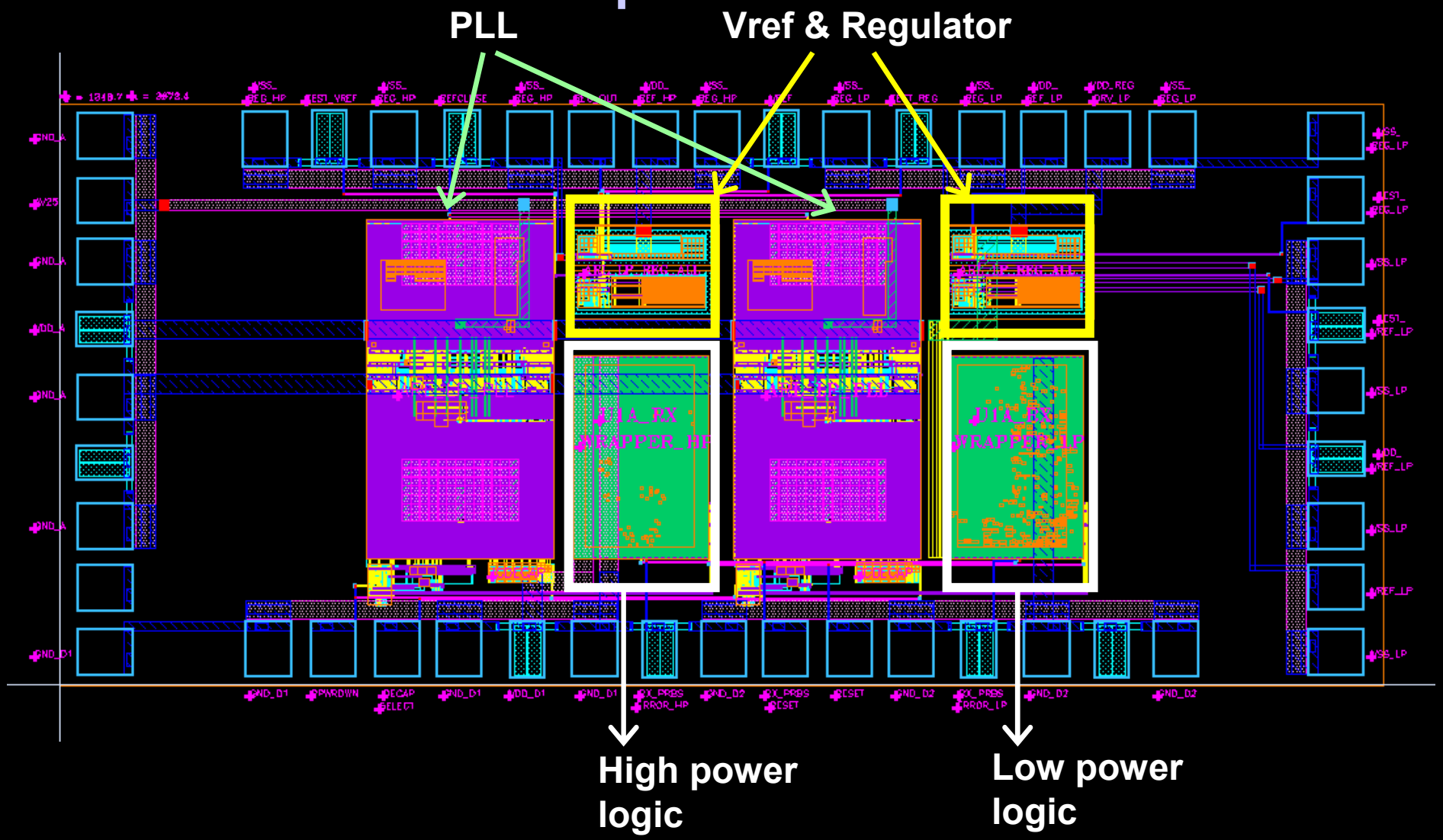


3. Regulation Improves Robustness

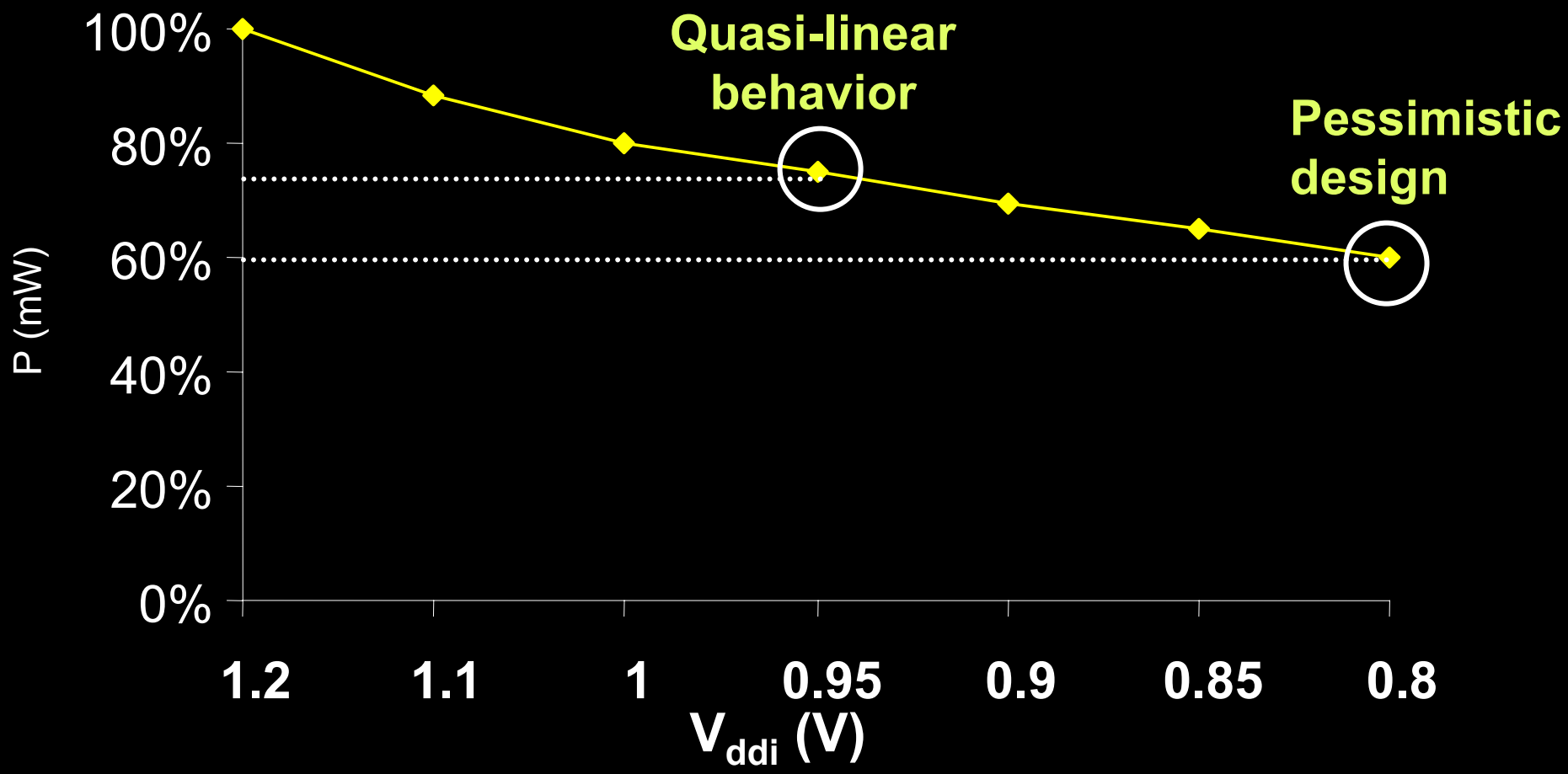
- **Reduced impact of supply variation**
  - Smaller effective corners
  - Selectable voltage



3. 3.2 Gbit/s 130nm Chip



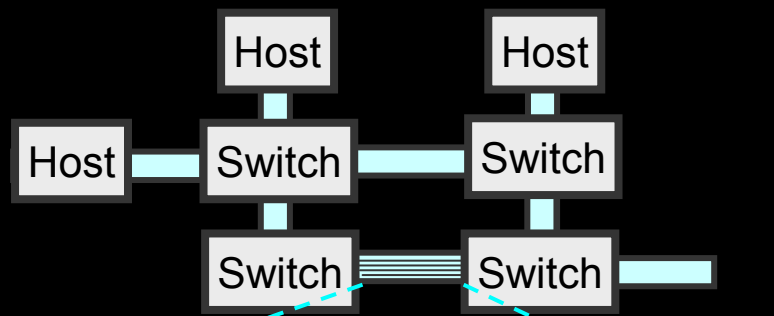
3. Observed Power Savings

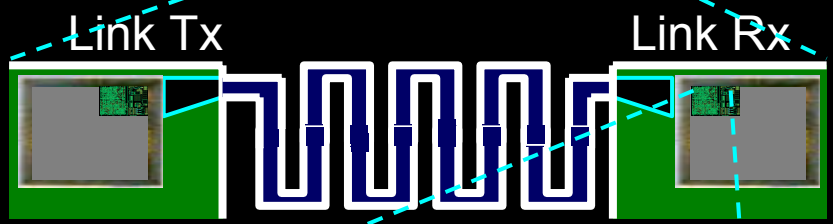


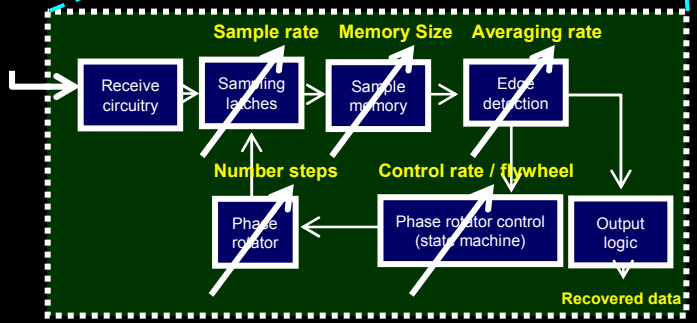
3. Advantages of Semi-Custom Voltage Islands

- **Adaptability**
  - To voltage supply variations, to manufacturing variations
  - Supply is digitally selectable and accurately regulated
- **Productivity**
  - Selective custom design helps design convergence (25%)
  - Logic can also be selectively shifted to high supply island
- **Low overhead**
  - Custom design may even reduce area!
  - No impact on system supply distribution

4. Productive Design of Adaptive Link Networks







Link Rx Circuit

Trade-off energy-bandwidth
Determine network

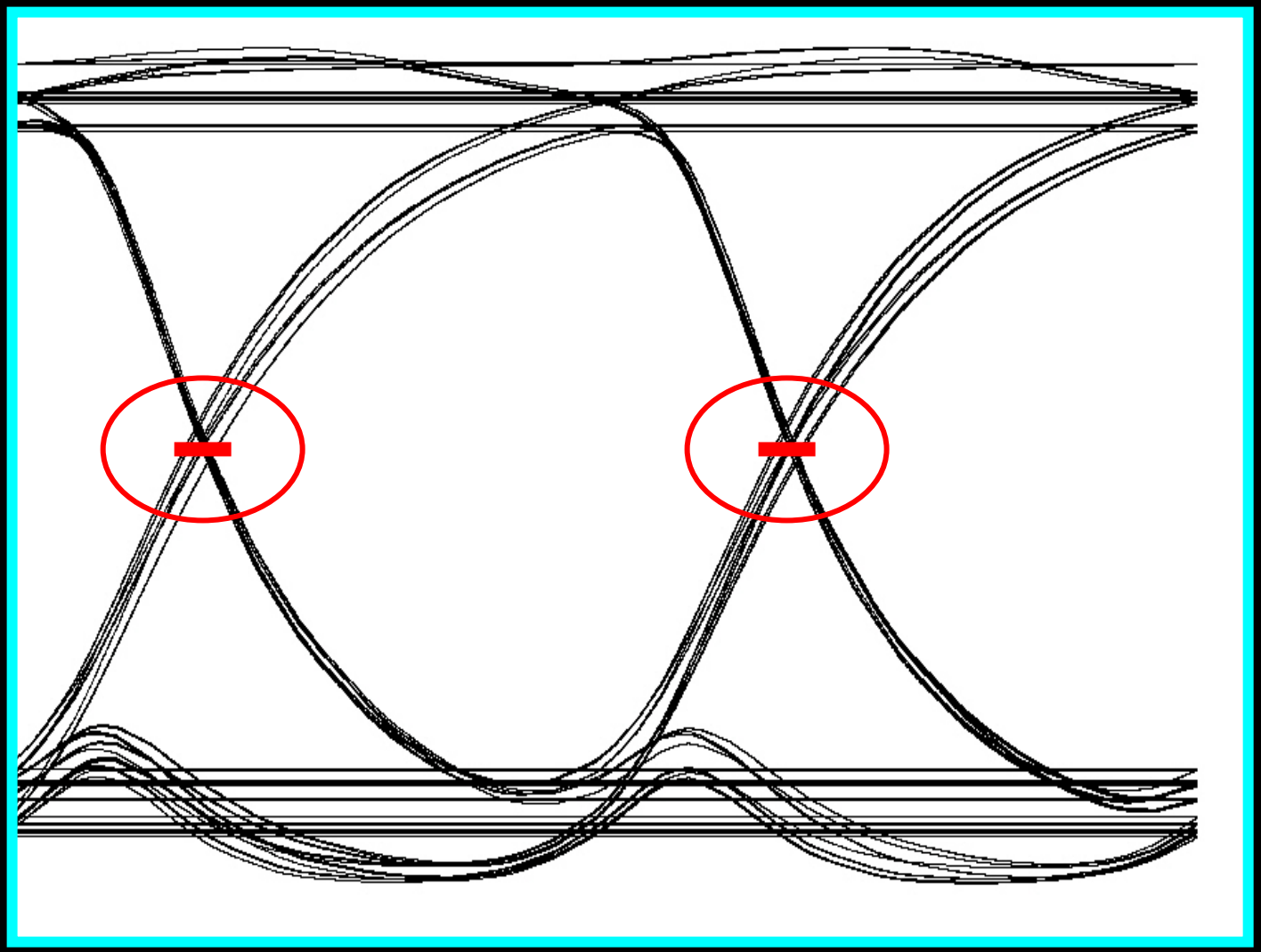
Trade-off power-BER
Determine architecture modes

Trade-off power-jitter
Determine circuits

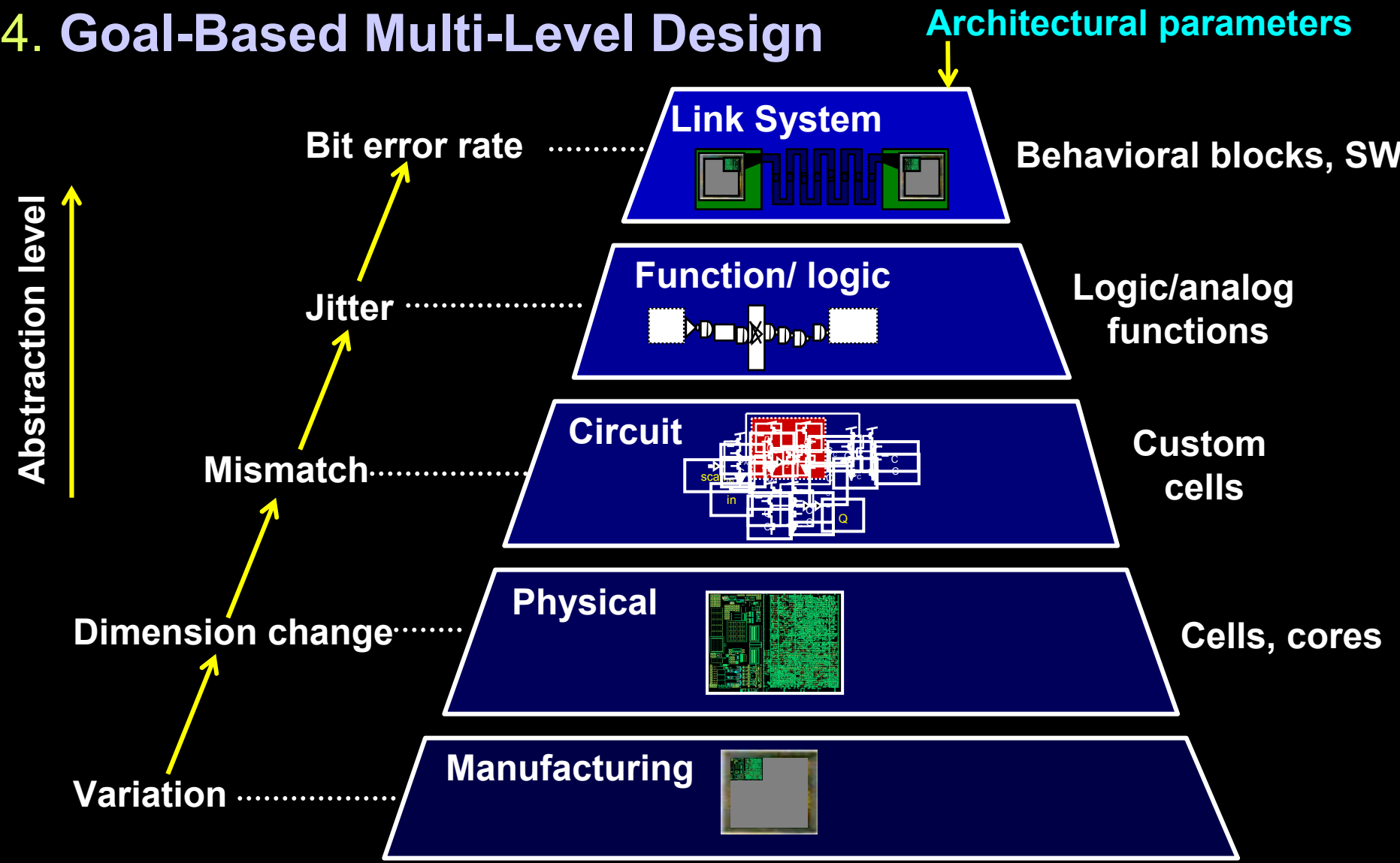
4. Multi-Level Design

- **Goals**
  - Adaptability → Flexible architecture definition
  - Productivity → Fast yet accurate exploration
  - Performance/power → Trade-off definition
- **Approach**
  - Relate BER performance to jitter and then to technology
  - Enable architecture to be parametrically varied
  - Allow explicit power-BER-BW goals and trade-offs

4. Jitter (Eye Closure) Determines Performance



4. Goal-Based Multi-Level Design



4. Types of Jitter

- **Random jitter**
  - Noise associated to devices (e.g., thermal transistor noise).
  - Phase-Locked-Loops tend to concentrate most of this jitter
- **Deterministic jitter**
  - Algorithmic and bandwidth limitations
    - Signal processing algorithm functionality (e.g. CDR filter)
    - Bandwidth limitations of analog circuits
  - Device mismatch and supply voltage variation
    - Related to technology → affected by process variations

4. Technology Versus Jitter

- **BER can be approximated as function of jitter (J)**

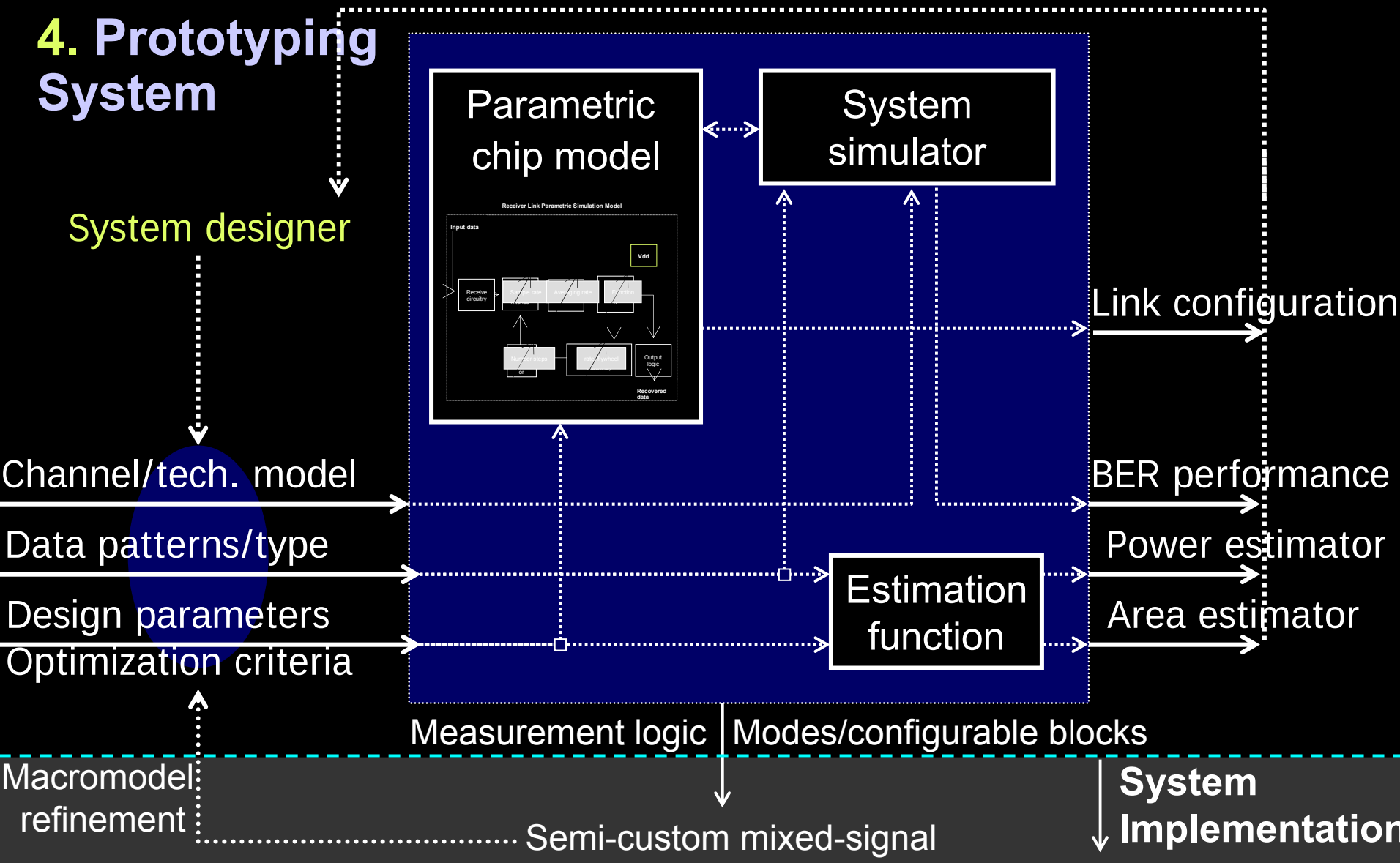
$$BER(J) \approx \int_{v}^{\infty} \frac{1}{\sqrt{2\pi(KJ)^2}} e^{\left(\frac{-x^2}{2(KJ)^2}\right)} dx$$

- **Jitter can be approximated as function of variability (σ)**

$$J \approx D_J + R_J \approx \sigma_V^2 + \sigma_M^2 + \sigma_N^2 + \sigma_B^2$$

Supply variation | **Device/wire mismatch** | **Random/ noise** | **Algorithmic/ bandwidth**

4. Prototyping System



Summary → Multi-Level Communications Design

Level	Design strategy	Adaptability	Productivity
Link design	Self-adaptive links	Application, environment 50% better power	Single design
	On-core problem determination	Post-manufacturing environment	Debugging
Circuit design	1/2-custom islands	Supply variations 25% better power	Design convergence
	Custom digital Custom analog	Manufacturing variations 10% better power	Design convergence

Summary

- **Future large computing systems require**
  - Adaptability, productivity, performance
- **Networking / communications key to these systems**
  - And power-efficiency as important as performance
- **Productive, energy-efficient, adaptive networks are possible**
  - Adaptive communications architectures
  - Intelligent on-core problem determination
  - Semi-custom multiple-voltage-domain link cores
  - Multi-level design methodology