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

- Almaden
- Watson
- Zurich
- Beijing
- California
- New York
- Switzerland
- China
- Japan
- Austin
- Haifa
- Delhi
- Tokyo
IBM Research – Communications Technology

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

Programming & user interface
- System software
- Architecture
- Technology
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 \( \text{mW} \)

Circuit design \( \mu W \)

Link Tx

Host

Switch

Host

Switch

Link Rx

Receive circuitry → Sampling latches → Sample memory → Edge detection

Number steps → Control rate / flywheel → Output logic

Recovered data

Sample rate → Memory Size → Averaging rate

Link Rx Circuit
Designing High-Bandwidth Links is Difficult

Fiber Channel, Optics, 1 Km

InfiniBand, Copper, 1 meter

Under-80mW Gbit/s serial link core

- **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

<table>
<thead>
<tr>
<th>Level</th>
<th>Design strategy</th>
<th>Adaptability</th>
<th>Productivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link design</td>
<td>Self-adaptive links</td>
<td>Application, environment 50% better power</td>
<td>Single design</td>
</tr>
<tr>
<td></td>
<td>On-core problem determination</td>
<td>Post-manufacturing environment</td>
<td>Debugging</td>
</tr>
<tr>
<td>Circuit design</td>
<td>1/2-custom islands</td>
<td>Supply variations 25% better power</td>
<td>Design convergence</td>
</tr>
<tr>
<td></td>
<td>Custom digital</td>
<td>Manufacturing variations 10% better power</td>
<td>Design convergence</td>
</tr>
<tr>
<td></td>
<td>Custom analog</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. Adaptive Communications Links

- Application requirements, environment impact performance
  Workload, frequency, quality of channel, package, chip layout

```
<table>
<thead>
<tr>
<th>Application</th>
<th>Margin for required BER</th>
</tr>
</thead>
<tbody>
<tr>
<td>BER=10-17</td>
<td>&quot;Easy&quot;</td>
</tr>
<tr>
<td>BER=10-15</td>
<td>&quot;Difficult&quot;</td>
</tr>
<tr>
<td>BER = 10-12</td>
<td></td>
</tr>
</tbody>
</table>
```

Application
1. “Difficult” Versus “Easy” Requirements

<table>
<thead>
<tr>
<th></th>
<th>Difficult</th>
<th>Easy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Workload</strong></td>
<td>Fast-varying frequency pattern</td>
<td>Uniform-frequency pattern</td>
</tr>
<tr>
<td><strong>Frequency</strong></td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td><strong>Channel</strong></td>
<td>Long across boards</td>
<td>Short chip-to-chip</td>
</tr>
<tr>
<td><strong>Package</strong></td>
<td>Low-cost low-yield</td>
<td>Ceramic, high quality</td>
</tr>
<tr>
<td><strong>Chip layout</strong></td>
<td>Longer wires</td>
<td>Shorter wires</td>
</tr>
</tbody>
</table>
1. Requirement-Based Design of Adaptive Links

Requirements space
- Requirement measurement blocks
- High-frequency jitter

Mapping
- Power modes

Link Design space
- Configurable link CDR blocks
1. Jitter (Eye Closure/Movement) Determines Difficulty
1. Adaptive Link Receiver

**Receiver**

- **Environment quality measurement**
  - Loop stats generator

**Complexity, V, frequency control**

**Regulator**

**Receive circuitry**

- **Data In** (from channel)
- **Vdd**
- **Frequency**

**Oversampling**

- **Sample memory**

**Edge detection**

**Phase generation**

- **Phase control** (state machine)

**Data select + output**

**Data out** (in chip)
1. Example: Adaptive Loop Latency

- **Low design overhead**
  Simple clock gating + control logic

![Diagram of phase control state machine]

- **Clk**
- **/n**
- **/2**
- **Edge detection**
- **Phase control state machine**
- **Phase state**
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

- Environment quality measurement
  - Loop stats generator
- Test advisor engine
- BER checker
- Data select + output
- Result
  - Data out (in chip)
- Data In (from channel)
- Receive circuitry

- Oversampling
- Sample memory
- Edge detection
- Phase control (state machine)
- Phase generation

- Complexity, V, frequency control
- Regulator
2. Test Advisor Engine

Statistic₁ … Statisticₙ

PRBS result

PRBS conditions

rules

state

Issue determination table

Next step table

BER issue

Link/channel cause

Correction (optional)

Next test
## 2. Test Advisor Engine (Example)

<table>
<thead>
<tr>
<th>Pattern test results</th>
<th>Internal link stats</th>
<th>Outputs of advisor engine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Max. BER</td>
<td>HF Jitter (%UI)</td>
</tr>
<tr>
<td>PRBS 7bit</td>
<td>10^{-12}</td>
<td>&lt;50</td>
</tr>
<tr>
<td>PRBS 31bit</td>
<td>10^{-12}</td>
<td>&lt;50</td>
</tr>
<tr>
<td>PRBS 31bit</td>
<td>10^{-10}</td>
<td>&gt;75</td>
</tr>
<tr>
<td>PRBS 31bit</td>
<td>10^{-10}</td>
<td>&lt;50</td>
</tr>
<tr>
<td>PRBS 31bit</td>
<td>10^{-10}</td>
<td>&lt;50</td>
</tr>
<tr>
<td>JTPAT</td>
<td>10^{-8}</td>
<td>&lt;50</td>
</tr>
<tr>
<td>JTPAT</td>
<td>10^{-12}</td>
<td>&lt;50</td>
</tr>
</tbody>
</table>
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

- Receiver
  - Receiver/Sampling
  - Clock generation

- Retiming
- Critical logic
- Voltage regulator
  - LSSD shifters

- Data and clock extraction
- Clock control
- Parallel interface

Medium data → Recovered data

$V_{dde}$ (shareable)
3. Selective Custom Design Increases Flexibility

Vddi → Edge correlation path

Latch

AO22

Latch

LSSD level shift

Manual optimization

Multi-Vt merged logic

Latch

AO22

Latch

scan_in

in

Q

Qn

P

Pn

C

Cn

clk

clk

clk
3. Scannable Shifter+Latch

LEGEND

<table>
<thead>
<tr>
<th>$C_A$</th>
<th>Clock A</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_B$</td>
<td>Clock B</td>
</tr>
<tr>
<td>$C_C$</td>
<td>Clock C</td>
</tr>
<tr>
<td>in</td>
<td>Input</td>
</tr>
<tr>
<td>scan$_{in}$</td>
<td>Scan input</td>
</tr>
<tr>
<td>Q</td>
<td>Output</td>
</tr>
</tbody>
</table>
3. Integration/Customization Saves Power

- Conventional approach
- Scannable level-shifting latch
3. Using Critical Paths to Choose Supply Voltage

- Delay penalty
- Power savings

"Optimal" Supply region
3. Flexible Design Methodology

Custom digital

- Circuit
- Layout
- Extraction
- Characterization

Logic design
- Synthesis
- Timing
- Place+Route
- Extraction

Analog

- Circuit
- Layout
- Extraction

Verification
- Integration
3. Regulation Improves Robustness

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

![Graph](image)(Effective supply Impact (V))

- Non-regulated
- Regulated
3. 3.2 Gbit/s 130nm Chip

- PLL
- Vref & Regulator

- High power logic
- Low power logic
3. Observed Power Savings

- Quasi-linear behavior
- Pessimistic design
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

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

Architectural parameters

Bit error rate

Jitter

Mismatch

Dimension change

Variation

Manufacturing

Physical

Circuit

Logic/analog functions

Custom cells

Cells, cores

Behavioral blocks, SW

Link System

Function/logic

Abstraction level

Communications Research
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 \( J \)

\[
\text{BER}(J) \approx \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(KJ)^2}} e^{-\frac{x^2}{2(KJ)^2}} \, dx
\]

- Jitter can be approximated as function of variability \( \sigma \)

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

- Supply variation
- Device/wire mismatch
- Random/noise
- Algorithmic/bandwidth
4. Prototyping System

System designer

- Channel/tech. model
- Data patterns/type
- Design parameters
- Optimization criteria

Parametric chip model

System simulator

Estimation function

Measurement logic

Modes/configurable blocks

Semi-custom mixed-signal

Link configuration

BER performance

Power estimator

Area estimator

System Implementation

Macromodel refinement
# Summary → Multi-Level Communications Design

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