Recent advances in VLSI technology have created an increasing interest within the computer architecture community to build a new kind of “general purpose” processor that is able to run a broad class of applications including primarily those from the domain of embedded systems—graphics, wireless processing, networking, and various forms of signal processing. The interest in new architectures is compounded by a growing wire delay concern which limits the distance that information can travel in a single clock cycle. The realities of interconnect delay—and power consumption—seriously challenge the ability of microprocessor designers to fulfill the promise of Moore’s Law. As a result, new architecture designs are largely centered around scalable and distributed alternatives to current centralized microprocessor designs.

Several projects such as VIRAM [2] at Berkeley, Smart Memories [4] at Stanford, TRIPS [5] at UT-Austin, Raw [8] and SCALE [3] at MIT, and industrial efforts such as the Tarantula [1] extension to Alpha, have proposed architectures that organize silicon resources more effectively and as tiled-processors that are easily scalable. The DARPA program in Polymorphic Computing Architectures is also a research thrust in this new area, and emerging “polymorphic” architectures will eventually compete with traditional desktop processors (e.g., Pentium IV) not so much in better performance on desktop workloads, but in versatility, or the ability to run a broader class of applications more effectively. We also expect that architectures that are more versatile are also likely to run complex real-world applications more effectively, since complex applications are often comprised of diverse components. One such versatile, tiled-processor architecture (TPA), is the Raw microprocessor which was designed and implemented at MIT.

Raw divides the chip into a two-dimensional mesh of sixteen programmable tiles, and interconnects them through on-chip, point-to-point scalar operand networks (SON) [7]. The Raw processor can issue sixteen different floating-point, integer, load, store, or branch instructions each cycle. It also has a large set of registers and a distributed memory hierarchy. The SON is exposed to the Raw compilation infrastructure which orchestrates the flow of data within the network for streaming computation and fine-grained instruction-level parallel-processing.

The focus on TPAs and architectural versatility necessitates new benchmark suites and metrics to accurately reflect the goals of the architecture community. Toward that end, we propose both a new benchmark suite—VersaBench—and a new metric called Versatility. VersaBench is a collection of applications from three central tiers—desktops, servers, and embedded systems—encompassing traditional integer workloads, floating-point and scientific applications, server computing, stream processing, and bit-level computation. VersaBench thereby attempts to better characterize the broad set of workloads that the new tiled-processor architectures are required to run.

The Versatility metric is inspired by SPEC rates [6]. For example, the SPEC CINT89 rate for an architecture is the geometric mean of the speedups of that architecture relative to a reference machine (specifically, the VAX 11/780)1 for each of the applications in the SPEC CINT89 suite. Computing the Versatility of an architecture is purposefully designed to mirror that of SPEC rates. Accordingly, like SPEC, Versatility takes the geometric mean of the speedups of an architecture for each of the applications in the VersaBench suite. Unlike SPEC rates however, the speedup of each application is not computed relative to a single reference machine, but rather relative to the architecture which provides the best performance for that application (in the 2004 time frame from known results at the time of this writing).

1The reference machines have changed over time. While the VAX 11/780 was the reference machine for SPEC CINT89 and SPEC CINT92, the SPARCstation 10/40 was the reference machine for SPEC CINT95, and the Sun Ultra5-10 workstation with a 300MHz SPARC processor is the reference machine for SPEC CINT2000.
**Versatile Tiled-Processor Architectures: The Raw Approach**

Table 1. Characteristics of the VersaBench workloads.

<table>
<thead>
<tr>
<th>benchmark category</th>
<th>data type</th>
<th>parallelism</th>
<th>control complexity</th>
<th>temporal locality</th>
<th>spatial locality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desktop Integer</td>
<td>integer</td>
<td>low</td>
<td>high</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>Desktop Float</td>
<td>float</td>
<td>medium</td>
<td>medium</td>
<td>medium to high</td>
<td>medium to low</td>
</tr>
<tr>
<td>Server</td>
<td>integer/float</td>
<td>high</td>
<td>medium to high</td>
<td>medium to high</td>
<td>medium to low</td>
</tr>
<tr>
<td>Embedded Stream</td>
<td>integer/float/bit</td>
<td>very high</td>
<td>low</td>
<td>low to high</td>
<td>very high</td>
</tr>
<tr>
<td>Embedded Bit</td>
<td>bit</td>
<td>very high</td>
<td>very low</td>
<td>very low</td>
<td>very high</td>
</tr>
</tbody>
</table>

Presentation Outline

The presentation will describe the Raw architecture, its implementation, and performance. We will focus on Raw’s ability to support a diverse set of applications (ranging from desktop to embedded workloads) and multiple forms of parallelism (including instruction-level-parallelism (ILP) for desktop applications, and stream parallelism for embedded computing) as represented by the VersaBench suite. We will also report detailed performance measurements that quantify the versatility of Raw compared to some widely deployed architectures. As a prelude, the measured versatility of the Raw processor is 0.7, while that of the Pentium III is 0.1. The Pentium’s relatively poor performance on stream benchmarks hurts its versatility. Although Raw’s versatility is better in comparison, the VersaBench suite highlights two clear areas that merit additional research. The first is in improving the architecture to better support embedded bit-level workloads: ASICs perform 2x-3x better than Raw. Another area of research focuses on desktop integer applications: Raw’s performance is 2x lower than a Pentium III for applications with low degrees of ILP.

VersaBench

The VersaBench suite consists of fifteen benchmarks that are grouped into five categories that span desktop, server, and embedded application workloads. The VersaBench applications are themselves drawn from various suites, and were selected because of the salient behavior and the properties they exhibit along various dimensions. For example, it is widely accepted that desktop applications have complex control structures, whereas streaming applications consist of relatively small computational kernels with simple control mechanisms. In all, we consider five property-dimensions when characterizing a benchmark; they are

— predominant data type: summarizes the predominant type-domain over which computation is performed,
— parallelism: quantifies maximum IPC (instructions per cycle) in a benchmark,
— control complexity: measures instruction temporal locality,
— data temporal locality
— data spatial locality

Intuitively, we believe the properties of each of the five benchmark-categories are as shown in Table 1. Accordingly, the VersaBench suite was created systematically by measuring the properties of numerous applications and selecting those that matched intuition. The presentation will include detailed results that map applications into the five-dimensional space.

Versatility Metric

We define the Versatility of an architecture as the geometric mean of the speedup of every application in the VersaBench suite relative to the architecture which provides the best performance for that application. Thus architectural versatility becomes quantitative, with ASICs (application specific integrated circuits) occupying the lowest end of the spectrum (i.e., Versatility(ASIC) = 0); as future process-technologies deliver higher clock frequencies, architectural versatility will increase beyond unity. Versatility measure servers to identify the areas where opportunities for architectural improvements are greatest, and those where further efforts will lead to marginal returns. For example, a new architecture that performs as well as a Pentium IV for desktop workloads has little to gain from further improvements targeted toward that application domain. In contrast, an architecture that fares poorly compared to the best streaming processor warrants attention that is focused on improving the performance of that architecture within the streaming context.
REFERENCES


Versatile Tiled-Processor Architectures
The Raw Approach

Rodric M. Rabbah
with Ian Bratt, Krste Asanovic,
Anant Agarwal
Processor Model

• Stable model for last few decades
  - Von Neumann architecture
  - Sequentially execute instructions
  - Simple abstraction
  - Easy to program
Change Is Around the Corner

- Processor performance not scaling as before
  - Wire delay and power

old view: chip looks small to a wire
- chip size
- distance signal can travel in 1 cycle

new view: chip looks much bigger to a wire, communication is expensive even on chip!

- How to effectively use transistors?
Many forward looking architectures are addressing the physical challenges

- MIT Raw processor
- MIT Scale processor
- Stanford Imagine processor
- Stanford Smart Memories processor
- UC David Synchroscalar
- UT Austin TRIPS processor
- Wisconsin ILDP architecture
- The original IBM BlueGene processor
Problems with Monolithic Designs

- Super-wide general purpose processors are no longer practical

- Centralized control with global operand routing

- Area, power, and frequency concerns
Spatial Architectures
Spatial Architectures

RF

Bypass Net

ALU

ALU

ALU

ALU

ALU

ALU

ALU

ALU

ALU

ALU
Spatial Architectures
Raw On-Chip Networks

- **2 Static Networks**
  - Software configurable crossbar
  - 3 cycle latency for nearest-neighbor ALU to ALU
  - Must know pattern at compile-time
  - Flow controlled

- **2 Dynamic Networks**
  - Header encodes destination
  - Fire and Forget
  - 15 cycle latency for nearest-neighbor
Distribute the Register File
Distribute the Rest
Tiled-Processor Architecture
Tiled-Processor Architecture

- Tile abstraction is quite powerful
  - e.g., power $\rightarrow$ resources used as necessary

- Easily scalable
  - All signals registered at tile boundaries, no global signals
    - Easier to Tune the Frequency
    - Easier to do the Physical Design
    - Easier to Verify

Make a tile as big as you can go in one clock cycle, and expose longer communication to the programmer
Close-up of a Single Raw Tile
The MIT Raw Processor

- 180 nm ASIC (IBM SA-27E)
- 16 tiles → 16 issue
- Core Frequency:
  - 425 MHz @ 1.8 V
  - 500 MHz @ 2.2 V
- Frequency competitive with IBM-implemented PowerPCs in same process
- 18 W (vpenta)
The Raw Goal

• Create an architecture that
  - Scales to 100’s-1000’s of functional units, memory ports
    • By exploiting custom-chip like features
      - Application-specific routing of operands

- Is “general purpose” (Versatile)
  • Run ILP sequential programs, scientific computations, server-style processing, streaming systems, and bit-level applications
  • Support standard General Purpose Abstractions
    - Context switching, caching and instruction virtualization
The New Performance Goal
Versatility

- Raw architecture as an “all-purpose” processor
  - Better SPECmark/Watt across the board
    • Higher SPECmark → think more MIPS compared to some reference machine (e.g., VAX 11/780)

Figure borrowed from DARPA PCA Forum
Application Domains

- 5 market-dominant application domains
  - Desktop Integer
  - Desktop Floating (Scientific codes)
  - Server (Throughput Based)
    - Ergonomic simulations, Grid computation, Transaction processing
  - Embedded Streaming
  - Embedded Bit-Level
How Applications Differ

- Desktop
  - Integer
  - Floating point (scientific)
- Server
- Desktop integer
- Bit-level
- Streaming
Distinguishing Application Domains

• Five *basis* properties
  - Data temporal locality
    • Quantify address reuse
  - Spatial temporal locality
    • Quantify address adjacency
  - Predominant data type
  - Parallelism
    • ILP, DLP, TLP, etc
  - Instruction temporal locality
    • Inverse of control complexity
Classifying Applications

• Quantitative metrics for the basis properties
  - Measure properties of different applications
• Cluster applications into domains
  - VersaBench

<table>
<thead>
<tr>
<th></th>
<th>Data Type</th>
<th>Parallelism</th>
<th>Instruction Temporal Locality</th>
<th>Data Temporal Locality</th>
<th>Data Spatial Locality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desktop INT</td>
<td>integer</td>
<td>low</td>
<td>low</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>Desktop FLT</td>
<td>float</td>
<td>medium</td>
<td>medium</td>
<td>medium</td>
<td>medium</td>
</tr>
<tr>
<td>Server</td>
<td>integer/float</td>
<td>high</td>
<td>low to medium</td>
<td>medium to high</td>
<td>low to medium</td>
</tr>
<tr>
<td>Streaming</td>
<td>integer/float</td>
<td>very high</td>
<td>high</td>
<td>low to high</td>
<td>very high</td>
</tr>
<tr>
<td>Bit-Level</td>
<td>bit</td>
<td>medium-high</td>
<td>very high</td>
<td>low</td>
<td>very high</td>
</tr>
</tbody>
</table>
VersaBench Status

• 15 total benchmarks
  - 3 per category
    • Drawn from SPEC INT/FP, Raw, StreamIt, DIS (AAEC), USC ISI
  - Manageable size, encourages evaluation using the entire suite
  - Available online at http://cag.csail.mit.edu/versabench

• Benchmarks selected systematically
  - MIT Technical Memo 646, June 2004
    Rabbah, Bratt, Asanovic, Agarwal
Proposed Metric: Versatility

Versatility (VersaBench)
Geometric Mean of Speedup relative to best performing machines

SPECmark (SPEC)
Geometric Mean of Speedup relative to a single reference machine

• Normalization to the best performing machines identifies areas for improvements
  - This is especially important → VersaGraphs
  - Not another mean over N benchmarks

• High Versatility mark implies architecture is good across the board
VersaGraph Example

speedup of an ideal machine

speedup relative to best machine

desktop desktop server stream bit-level
integer float
VersaGraph Example

speedup of a general purpose machine (e.g., P3)
VersaGraph Example

speedup of an ASIC

speedup relative to best machine

desktop desktop server stream bit-level integer float
VersaGraphs For Real Architectures

Also compared against Athlon 64 and Itanium 2

P4 (2.8 GHz)

P3 (600 MHz)

Raw (425 MHz)

integer
mcf, parser

scientific
bmm, vpenta

server
mgrid, dbms

stream
corner, radio

bit-level
80211a, 8b10b
Raw Homepage
http://cag.csail.mit.edu/raw
download papers, benchmarks, …