Adaptive Solutions

CNAPS

(Connected Network of Adaptive ProcessorS)

Dan Hammerstrom

Gary Tahara

Adaptive Solutions, Inc.

Adaptive Solutions

The CNAPS Architecture is

- highly parallel,
- highly integrated,
- relatively low cost,

Engine for

- the emulation of Artificial Neural Networks (ANN),
- pattern recognition,
- classification, and
- image processing.

Adaptive Solutions

CNAPS ARCHITECTURE GOALS

- FLEXIBILITY

  all ANN (Artificial Neural Network) algorithms and other non-ANN pre and post-processing functions
- ADAPTABILITY

  on chip learning
- COST EFFECTIVENESS

  inexpensive application deployment
- HIGH-SPEED

  for real-time applications
- EASE OF SYSTEMS INTEGRATION

  interface cleanly to existing digital systems
- SIMPLICITY

  less expensive implementation, easy to program

Adaptive Solutions

ALL DIGITAL CMOS IMPLEMENTATION

- Advantages over CMOS analog:
  - manufacturability
  - computational precision
  - flexibility in on-chip functionality
  - ease of system implementation
  - standard bus oriented interface
  - digital input/output formats
- Multiple chips can be arbitrarily combined to create large systems

Adaptive Solutions

PROGRAMMABILITY

- Increases range of algorithms
- complex algorithms like LVQ2 or self-organizing maps can be implemented entirely on chip
- True on-chip learning with arbitrary learning algorithms
- Many non-ANN functions
- Preprocessing: DFT, digital filtering, segmentation
- Postprocessing: rule based (fuzzy and non-fuzzy)
- Image processing
- data compression (e.g., JPEG and MPEG)

Adaptive Solutions

FIXED POINT PRECISION

- Conforms to idea that ANN consists of large numbers of simple, low precision processing elements
- Sufficient for neural network algorithms (Baker and Hammerstrom, Holt and Baker)
- Reduces silicon area requirements
- Simplifies design
- NetTalk, for example, showed no appreciable difference between 32 bit floating point and 16 bit integer
- PN size with reduced fixed point precision is small enough to effectively utilize silicon redundancy

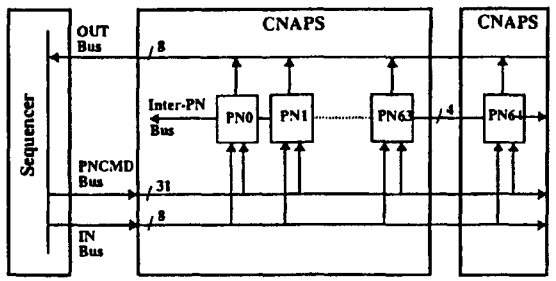
Adaptive Solutions

THE ARCHITECTURE

- A linear array of PNs
- SIMD control, each PN executes the same instruction each clock
- External sequencer with writeable program store
- Broadcast instruction bus (31 bits), PNCMD
- Broadcast input data bus (8 bits), IN bus
- Single, arbitrated output bus (8 bits), OUT bus

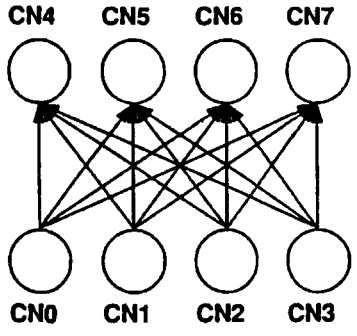
Adaptive Solutions

Basic CNAPS Array Layout:



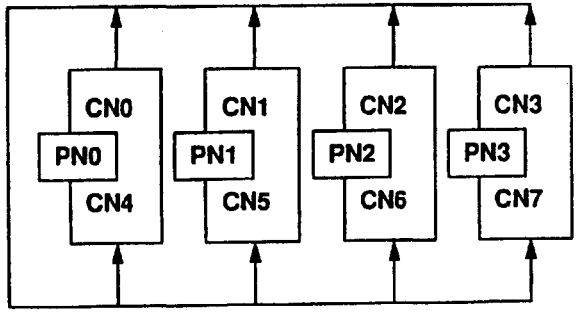
Adaptive Solutions

Feed forward network emulation:



Broadcast by PN0 of CN0's output to CN4, 5, 6, 7 takes 1 clock

N^2 connections in N clocks



Adaptive Solutions

PN ARCHITECTURE

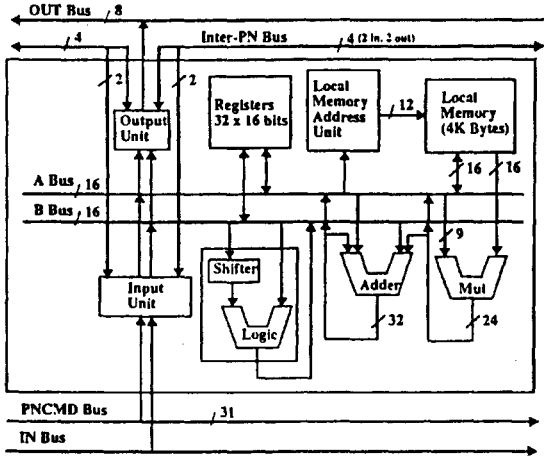
- Simple digital signal processor (DSP) like configuration
- Each PN has its own private memory
- Three basic weight modes: 1, 8, and 16 bits (1 bit mode computes 8 connections per clock)
- Two's complement representation
- External 8 bit busses, internal 16 bit busses (automatic assembly/disassembly)
- 9x16 multiplier, 16x16 multiply in two cycles
- 32 bit accumulation, direct path from multiplier to adder
- Logic/shift unit for binary point adjustment
- per PN conditional execution facility

PN ARCHITECTURE Cont'd

- Register file (32x16) for storing constants
- Saturation arithmetic: overflow saturates to maximum negative or positive number
- Zero bias bit truncation
- Separate weight address adder
- Maximization function to allow Winner-take-all networks
- Arbitrary max regions are possible
- Inter-PN bus for nearest neighbor (1-D) data transfer and OUT bus arbitration
- Support for sparse connection storage

Adaptive Solutions

PN Architecture:



Adaptive Solutions

PHYSICAL IMPLEMENTATION

- CNAPS-1064 implements PN array portion of CNAPS
- Each 1064 has 64 PNs with 4K bytes of weight storage per PN
- Total connections per chip:
  - 2M 1 bit weights
  - 256K 8 bit weights
  - 128K 16 bit weights
- At 25 MHz (all PNs busy):
  - 12.8 billion CPS for 1x1
  - 1.6 billion CPS for 8x8 or 8x16
  - 0.8 billion CPS for 16x16

Adaptive Solutions

CNAPS-1064

- 0.8 micron, 2 metal CMOS
- 1 inch on a side (2.54 cm)
- 11 million active transistors
- 50 square micron state of the art SRAM cell
- 80 PNs are fabricated on each die
- at 25 MHz, nominal power dissipation is 4 watts
- 106 pins
- proprietary redundancy technique switches out faulty PNs, significantly lowers "cost per PN"

Adaptive Solutions

LEARNING PERFORMANCE

Back-propagation learning speed, connections updated per second:

CNAPS (4 chips, 20 MHz):	**1067M**
Convex C1:	**.06M**
Cray 2:	**7M**
Connection Machine:	**40M**
Warp:	**20M**

Adaptive Solutions

SINGLE CHIP PERFORMANCE

NetTalk

- **Back-propagation learning**
- **203 input nodes**
- **64 hidden nodes**
- **26 output nodes**
- **76,700 training vectors (15 passes)**

The entire network fits onto a single CNAPS-1064 and learns the above vectors in

- **6 seconds**
- **183M CUPS**

Adaptive Solutions

SINGLE CHIP PERFORMANCE

LVQ2

- 960 reference vectors
- 256 inputs per vector (8 bits element)
- one pass through entire network (non-learning): 250 microseconds

Discrete Fourier Transform

- 128 point, real (1 chip): 21 microseconds
- 256 point, complex (8 chips): 42 microseconds

JPEG (image compression)

- image: 512x512, 3 colors, effectively 16 bits/pixell
- 41 frames/sec

Adaptive Solutions

OTHER APPLICATIONS

- Mitsubishi Kanji OCR:

  - 3500 characters (printed)
  - 15 fonts
  - 400 cps
  - > 99.5% accuracy

- Speech (1 second):

  100 ms - fourier spectral analysis, and to compute seven time functions, , including pitch, spectral derivatives and energy functions

  23 ms - to do neural network classification

Adaptive Solutions

CONCLUSIONS

We believe that we have attained our design goals:

- **this processor chip will provide, in a personal computer environment,**

  state of the art performance

  in a variety of image and signal processing applications,

  including neural network emulation

- **it is relatively easy to program and to integrate into a larger system**