CCIT Report #436
July 2003

# Parallel VLSI Architecture and Parallel Interleaver Design for Low-Latency MAP Turbo Decoders

Reuven Dobkin, Michael Peleg and Ran Ginosar
VLSI Systems Research Center, Electrical Engineering Department
Technion—Israel Institute of Technology
Haifa 32000, Israel
[ran@ee.technion.ac.il]

*Abstract* - Standard VLSI implementation of turbo decoding requires substantial memory and incurs a long latency, which cannot be tolerated in some applications. A novel parallel VLSI architecture for low-latency turbo decoding is described, comprising multiple SISO elements, operating jointly on one turbo coded block, and a new parallel interleaver. The design algorithm for the parallel interleaver is presented, enhancing the error correction performance of the parallel architecture. Latency is reduced up to twenty times and throughput for large blocks is increased up to five-fold relative to sequential decoders, using the same silicon area, and achieving very high coding gain. The parallel architecture scales favorably — latency and throughput improvement with growing block size and chip area.

Index Terms: maximum a posteriori (MAP) algorithm, turbo codes, parallel architecture, VLSI architecture, decoders, interleaver.

## 1. Introduction

Turbo-codes with performance near the Shannon capacity limit have received considerable attention since their introduction in 1993 [1][2]. Optimal implementation approaches of turbo codes are still of high interest, particularly since turbo codes have become a standard for 3G.

VLSI *sequential architectures* of turbo decoders consist of $M$ Soft-Input Soft-Output (SISO) decoders, either connected in a pipeline, or independently processing their own encoded blocks [3][4][5]. Both architectures process $M$ turbo blocks simultaneously and are equivalent in terms of coding gain, throughput, latency and complexity.

For the decoding of large block sizes, sequential architectures require large amount of memory per SISO for $M$ turbo blocks storage. Hence, enhancing throughput by duplicating SISOs is area inefficient. In addition, latency is high due to iterative decoding, making the sequential architecture unsuitable for latency-sensitive applications such as mobile communications, interactive video and telemedicine.

One way to lower latency is to reduce the number of required decoding iterations, but that may degrade the coding gain. An interesting tree-structured SISO approach [6] significantly reduces the latency, at the cost of an increased area requirement. Parallel decoding schemes [7][8] perform the SISO sliding window algorithm using a number of sub-block SISOs in parallel, each processing one of the sliding windows. Those