A 40-nm 144-mW VLSI Processor for Real-time 60-kWord Continuous Speech Recognition

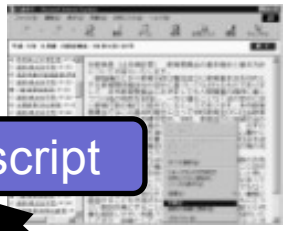
Guangji He, Takanobu Sugahara, Tsuyoshi Fujinaga, Yuki Miyamoto, Hiroki Noguchi, Shintaro Izumi, Hiroshi Kawaguchi, and Masahiko Yoshimoto.

Kobe University, Japan

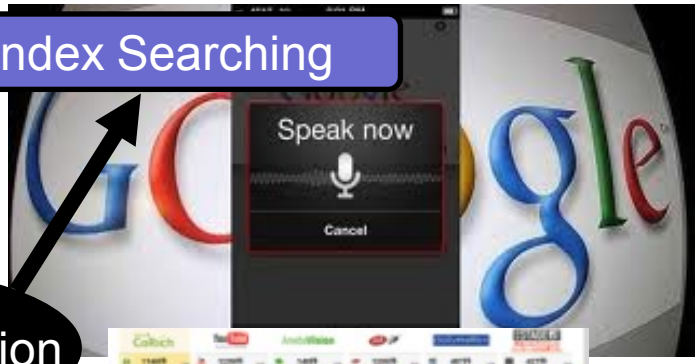
Jan. 23th, 2013, ASP-DAC

Background

Automatic Transcript

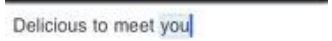


Index Searching



Speech Recognition

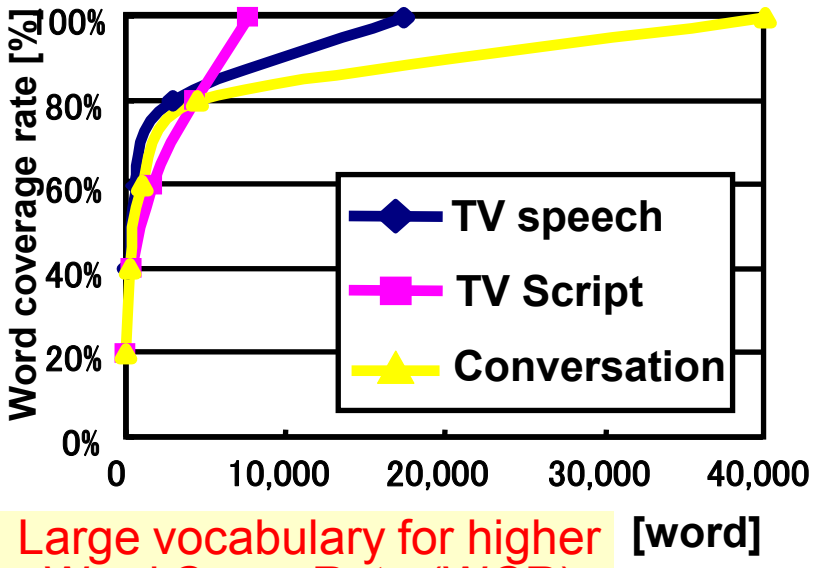
Voice Mail



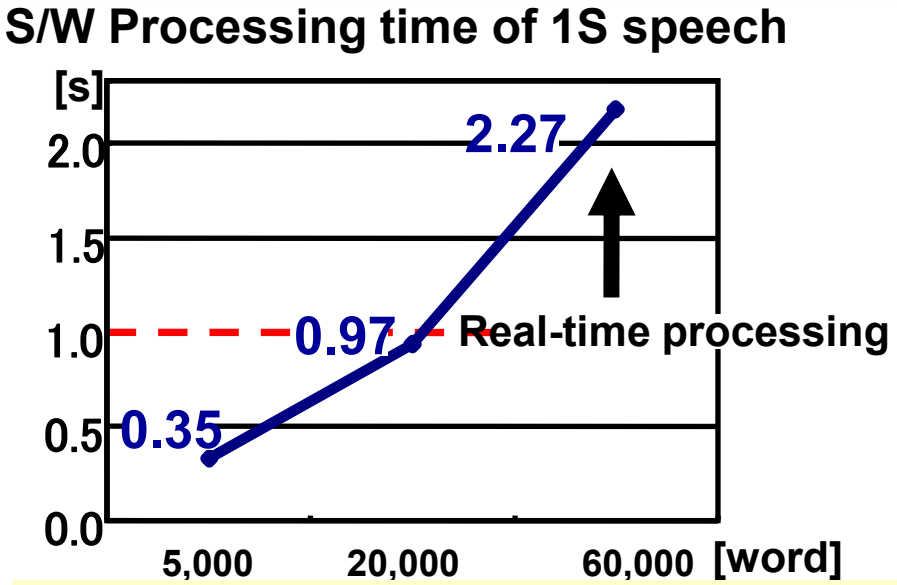
Human Interface

Speech Recogtnition is widely used in Mobile Systems, Ubiquitous Systems and Robotics

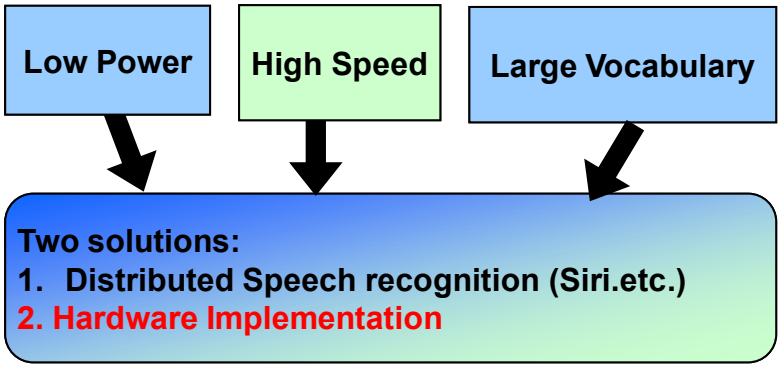
Background

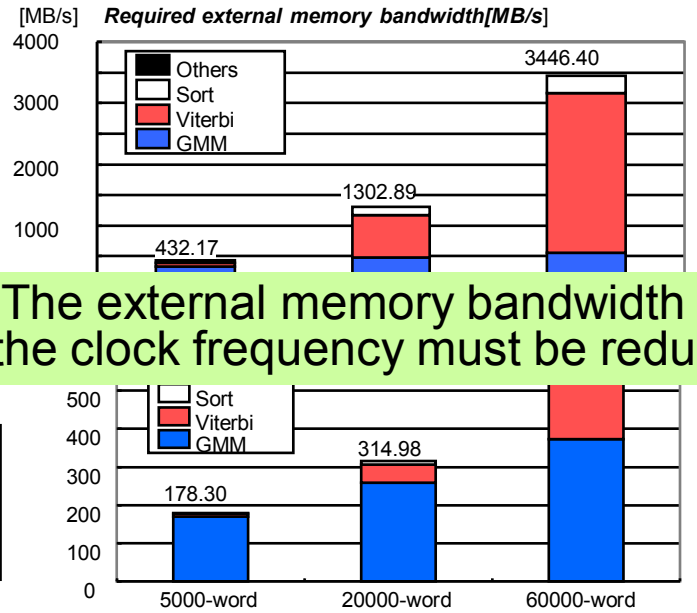


Large vocabulary for higher Word Cover Rate (WCR)



Limited speed and High power by S/W





The external memory bandwidth and the clock frequency must be reduced!

"2" is more stable than "1" because DSR systems may have long latency due to "Server Congestion" or " Signal interfere"

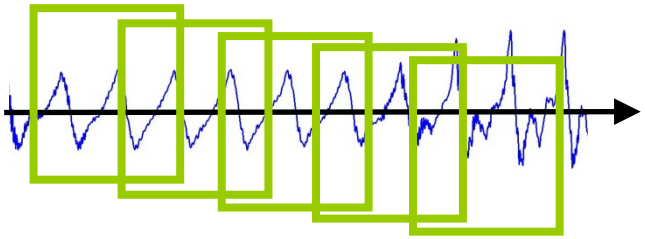
Algorithm optimization

Two-Stage language model searching

Reducing much computation and memory access with little accuracy degradation.

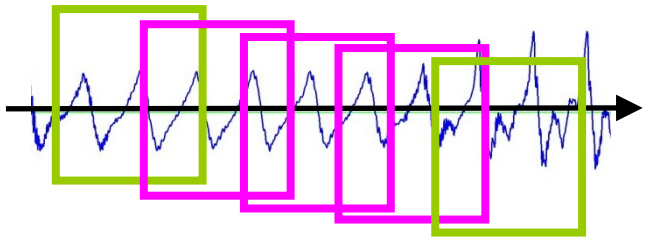
- Detailed language model search for all cross-word transitions
- Simplified language model search for the top 10 important transitions

Original algorithm:



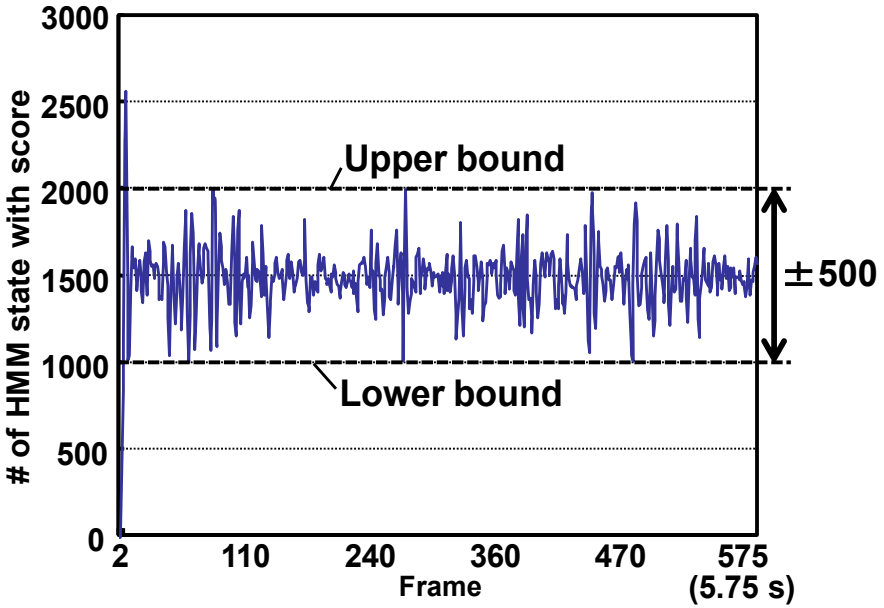
Detailed language model search for every frame

Proposed algorithm:

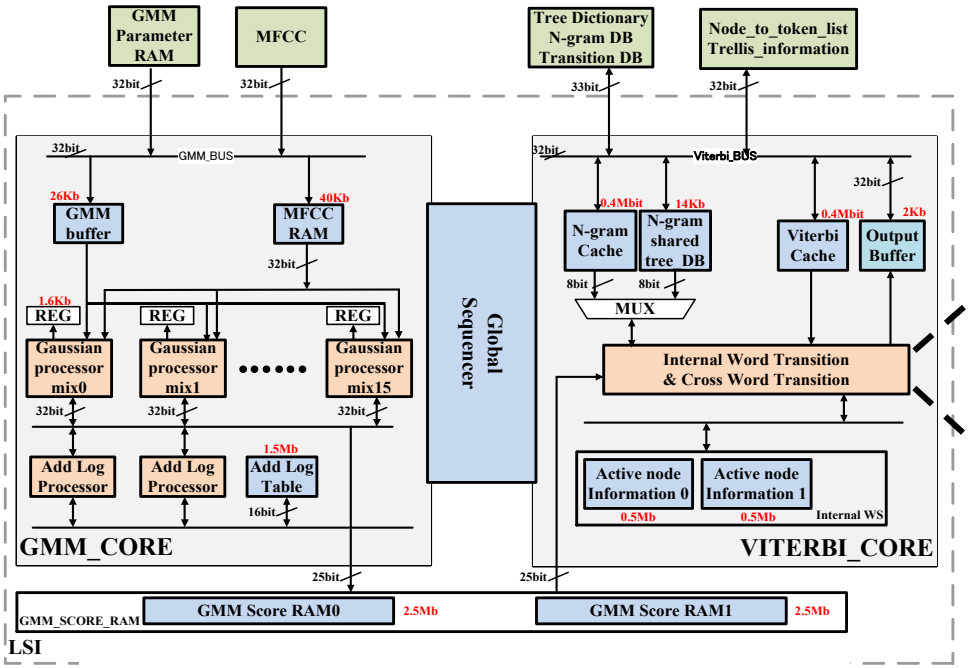


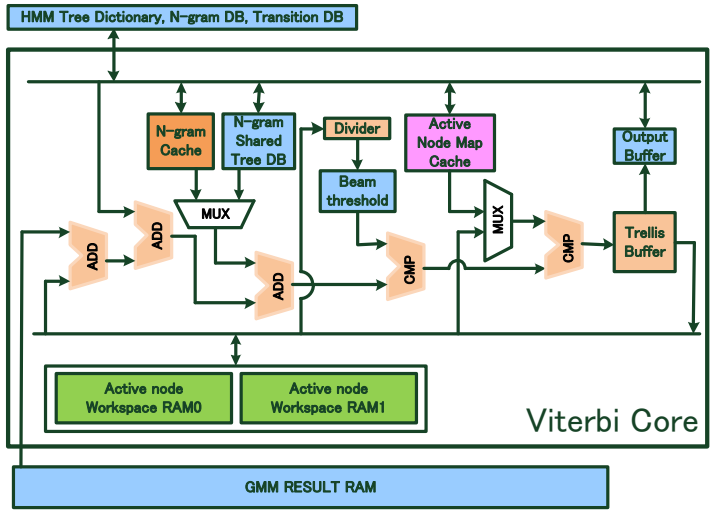
Detailed language model search for every n frame

Beam pruning using a dynamic threshold

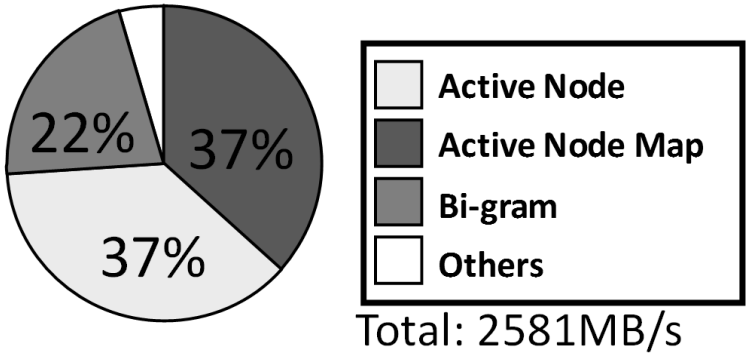


Cache Architecture

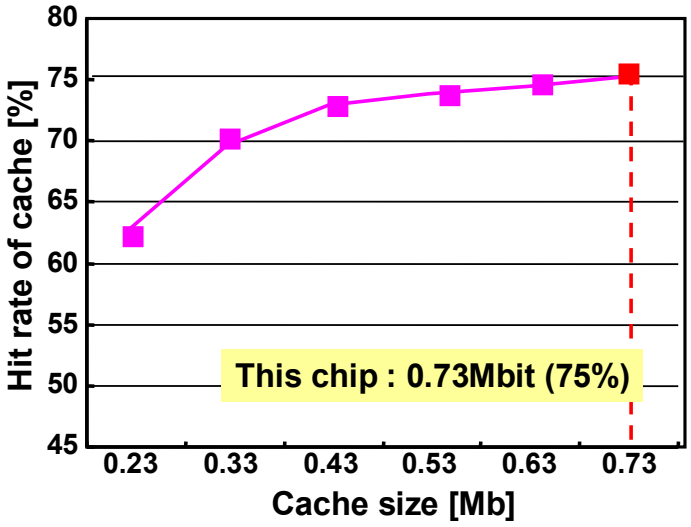




•Viterbi Cache architecture

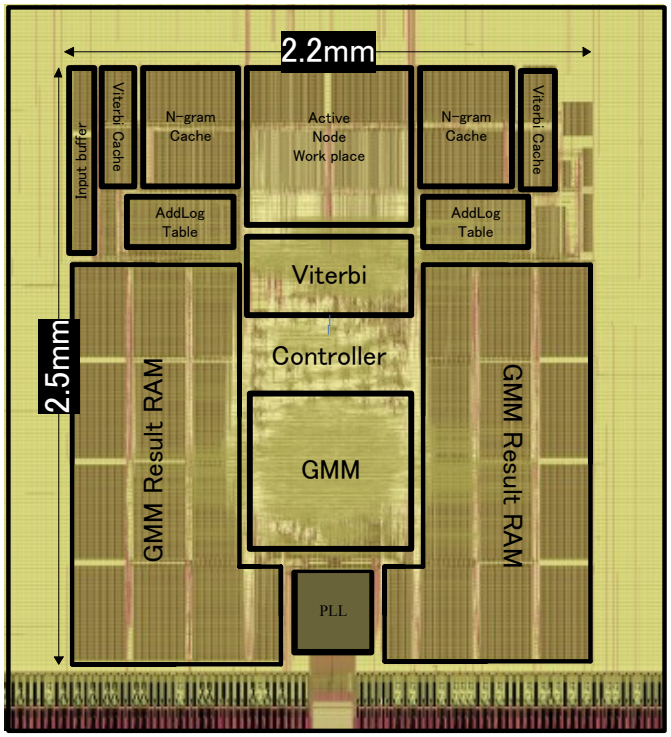


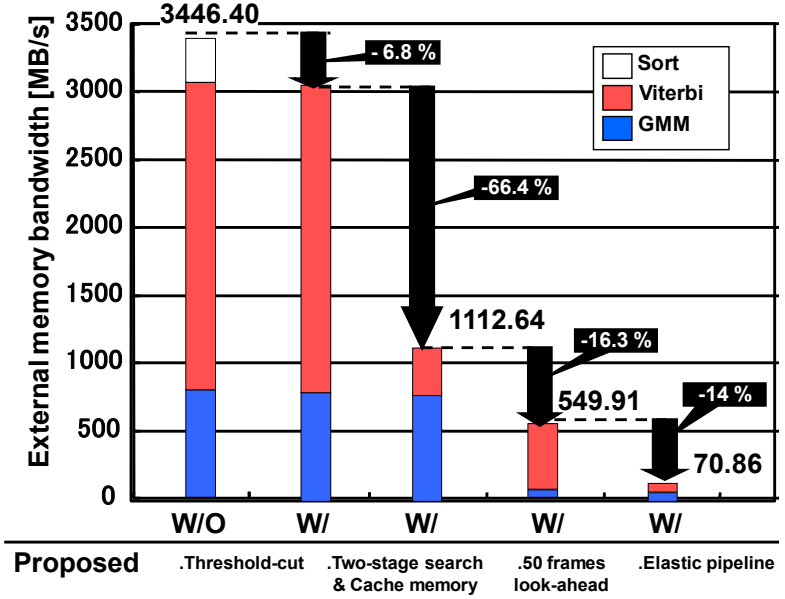
Total: 2581MB/s



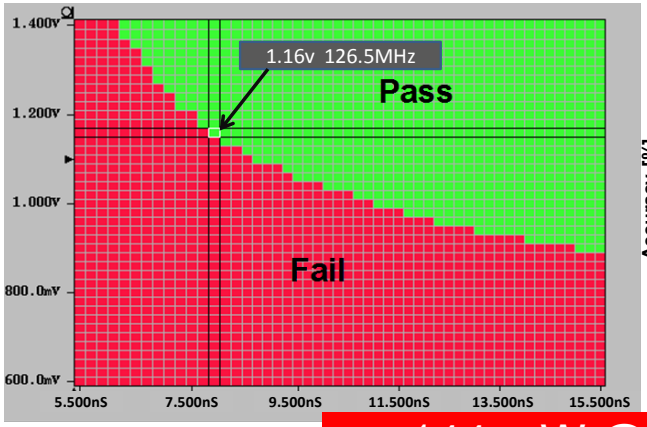
Specialized Cache architecture is proposed to improve the cache hit rate

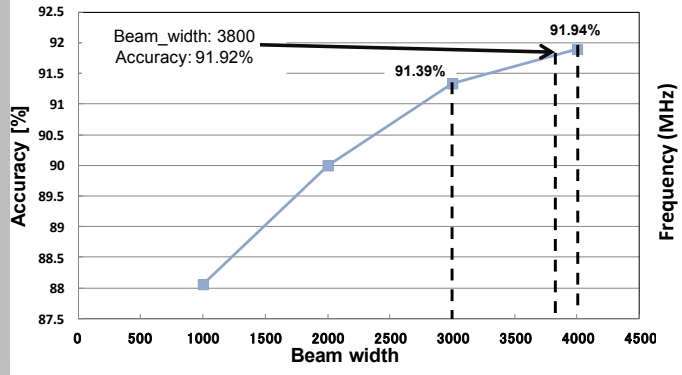
Measurement Results

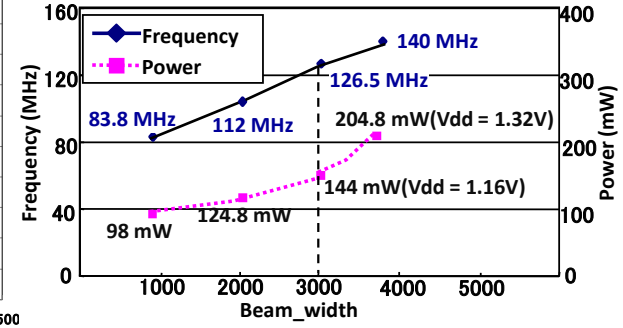




Achieves 95% external memory bandwidth reduction and 78% of the required frequency reduction







144 mW @ 126.5MHz for real-time processing