Parallel Layered Decoding Architecture

In conventional layered decoding algorithm (LDA), data processing of lower layer requires values of messages coming from upper layer, making the LDA totally serial. The decoder has to process data from the first layer until the last layer, which causes serious throughput limitation due to large decoding latency. Parallel layered decoding architecture (PLDA) is able to maintain high throughput using much less hardware resources. It is similar to the classic partly parallel decoding architecture, but with message passing between different layers and therefore can also accelerate the convergence speed. PLDA allows parallel processing among different layers while the updating of messages is in sequential. CTV messages \( r_m \) are calculated during the horizontal processing and then transmitted to overlapped positions in other layers. Specific network precisely regulates the transmission directions of the messages to guarantee that at another layer, messages from the previous layer must arrive before they are used.

Therefore, we can conclude the message passing rule between different layers at the vertical block \( j \) as follows: summation message at layer \( l \) will be transmitted to the layer that has the maximum permutation value cyclicly smaller than the permutation value at layer \( l' \).

Overall Decoder Architecture

The updated APP messages obtained by one layer are stored in the APP Mem Bank and used by another layer. Hence, convergence speed can be doubled and decoding latency per iteration can also be reduced significantly. Besides, a VNU and a CNU of the same layer can be merged and split into several pipelined stages by inserting registers. As a result, overall decoding frequency is expected to be enhanced. All of these methods result in great improvement in decoding throughput.

Result Highlights

- Decoding throughput: 2.2 Gbps
- Operating frequency: 950MHz
- Core area: 2.8mm²