# Energy-Efficient Soft-Output Trellis Decoder Design Using Trellis Quasi-Reduction and Importance-Aware Clock Skew Scheduling

Yang Liu, Fei Sun, and Tong Zhang
Electrical, Computer and Systems Engineering Department
Rensselaer Polytechnic Institute

***Abstract*— Energy-efficient implementation of high-speed soft-output trellis decoders is of great practical importance. This paper first presents an algorithm-level technique, referred to as quasi-reduced-state trellis decoding, that enables the use of reduced-state trellis decoding concept to reduce the energy consumption of decoding data storage without incurring any speed penalty. Then we propose to integrate this algorithm-level technique with an importance-aware clock skew scheduling approach that enables the use of aggressive voltage overscaling in decoding computation datapath at the cost of small decoding performance degradation. The integration of these two techniques can provide a wide and flexible design space to explore the decoding performance vs. decoding energy consumption trade-off for very high-speed soft-output trellis decoder implementations. The effectiveness has been demonstrated through 1Gbps soft-output Viterbi algorithm (SOVA) decoder ASIC design at 65nm technology node.**

## I. Introduction

High-speed soft-output trellis decoding is one of the key functions in modern data communication and storage systems. Well-known soft-output trellis decoding algorithms include BCJR algorithm [1] and soft-output Viterbi algorithm (SOVA) [2] and their variants. Due to their computation and data storage intensive nature, soft-output trellis decoders tend to be energy hungry, hence their energy-efficient implementation is of great practical importance.

At the algorithm level, energy reduction may be realized through reduced-state trellis decoding [3], [4] by merging a certain number of states in the original trellis into one super state. Decision feedback is typically used to partially compensate the performance degradation incurred by trellis reduction. However, the use of decision feedback may degrade the achievable speed performance from two perspectives: (i) decision feedback itself may appear on the circuit critical path, which will directly degrade the speed performance, and (ii) more importantly, decision feedback prevents the use of some well-proven high-speed design techniques developed for full-state trellis decoding such as bit-level pipelining [5] and CSA transformation [6], [7].

In this work, we propose a technique to eliminate the decision-feedback-induced speed bottleneck in reduced-state decoders while largely maintaining the power saving potential. Its basic idea can be described as follows. A soft-output trellis decoder consists of two main functional blocks including state metric computation and data storage. The use of decision feedback in conventional reduced-state decoders is due to the reduced-state metric computation. Meanwhile, as demonstrated in prior work (e.g., [7]), the energy consumption of the data storage block tends to be much higher than that of the state metric computation. Therefore, if we map the data storage onto a reduced-state trellis and map the state metric computation onto the original full-state trellis, the decision feedback and its resulted speed bottlenecks will be completely eliminated while the energy saving may be largely maintained. Accordingly, we call such a scheme as *quasi-reduced-state decoding* due to the fact that the reduced-size trellis structure is only used for the data storage. This is in sharp contrast to all the prior work that assume both metric computation and data storage always operate on the same reduced-size trellis.

As the trellis reduction factor of quasi-reduced-state trellis decoders increases, the state metric computation will be responsible to a higher percentage of the overall decoder energy consumption. Hence, it is desirable to further reduce the energy consumption of the state metric computation. Voltage scaling is an effective circuit-level technique to reduce the energy consumption. In conventional practice, voltage scaling is lower bounded by $V_{\text{dd-crit}}$ under which the critical path delay equals the desired clock period. Voltage oversclaing (VOS) (i.e., overscale the supply voltage below $V_{\text{dd-crit}}$) results in the risk of transient timing errors. Motivated by the observation that transient timing errors on different signals in state metrics lead to largely different decoding performance degradation, we can apply clock skew scheduling, a well-known circuit design technique to adjust the circuit timing slack, in such an importance-aware manner that those more important paths have larger timing slacks and hence are more immune to VOS-induced errors. Such importance-aware clock skew scheduling, as first proposed in [8], can push the energy efficiency envelope of the state metric computation at minimal decoding performance degradation.

In summary, this paper presents a new quasi-reduced-state trellis decoding strategy to reduce the data storage energy consumption in soft-output trellis decoders at no cost of speed. This technique can be directly integrated with an importance-aware clock skew scheduling technique [8] that enables the use of voltage overscaling in trellis state metric computation. These two techniques together enable a wide spectrum of energy vs. performance trade-offs in soft-output trellis de-

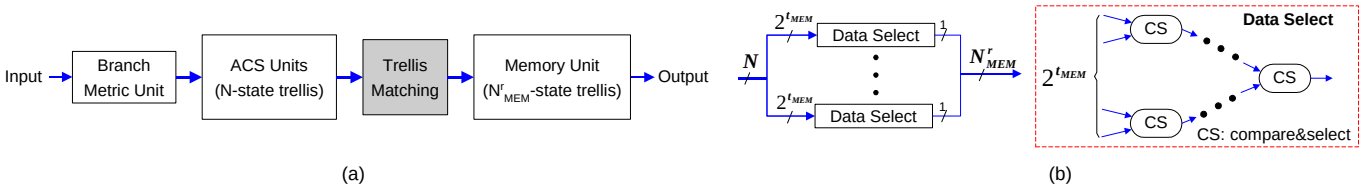


Fig. 2. (a) The general structure of the proposed quasi-reduced-state trellis decoder. The original $N$-state trellis is reduced to an $N^r_{MEM}$-state trellis ($N = 2^{t_{MEM}} N^r_{MEM}$) for implementing the memory unit, and (b) the structure of the trellis matching block.

coder design. For the purpose of demonstration, we use the SOVA decoder design for partial response channel in magnetic recording as a test vehicle. The corresponding full-state trellis has 16 states and various configurations/combinations of the presented two techniques are considered. The decoders are designed using the Synopsys tool set with a 65nm CMOS standard cell library and operate at 1Gbps decoding throughput. The energy saving and decoding performance in presence of VOS-induced errors are obtained through post-synthesis power estimation and circuits simulations.

## II. Quasi-Reduced-State Decoder Design

The fundamental approach of all the prior work on reduced-state trellis decoder design can be summarized as follows: Given the trellis reduction factor $t$ that is a positive integer, we group each $2^t$ states in the original $N$-state trellis into a single state, hence the number of trellis states reduces to $N^r = N/2^t$. However, if we directly map the chosen decoding algorithm onto this new $N^r$-state trellis, it will result in a significant decoding performance degradation compared with the full-state decoding. A $t$-depth decision feedback is typically used to reduce such performance degradation, where the basic idea is to use the immediate tentative decisions of the previous $t$ bits along the survivor sequences to improve the accuracy of branch metric computation.

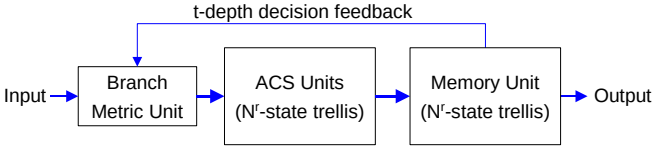


Fig. 1. The general structure of a reduced-state soft-output trellis decoder, where the original $N$-state trellis is reduced to an $N^r$-state trellis and $N = N^r \cdot 2^t$.

Fig. 1 illustrates the general structure of a reduced-state trellis decoder. We note that such conventional design approach has the following features: (i) In general, the decoding performance degradation will increase as $t$ increases, which directly restricts how much the trellis and hence the energy consumption can be reduced in practice; (ii) The decision feedback loop involves a relatively long propagation delay and constitute the essential critical path; (iii) The existence of decision feedback loop prevents the use of some well-proven high-speed design techniques for conventional full-state trellis decoder such as the bit-level pipelining and CSA transformation.

The reason for the last feature above can be briefly explained as follows. In those high-speed design techniques including bit-level pipelining and CSA transformation, the state metric computation and generation of the corresponding decision bit are *skewed* along the time axis in order to reduce the circuit implementation critical path. As a result, at each trellis depth, the computation for the decision bit and all the bits in the state metric are not finished at the same time. Therefore, when the branch metric is being computed, the required immediate tentative decisions of certain previous bits are not available yet, which makes it impossible to apply the decision feedback as in the conventional reduced-state decoder. Equivalently, the use of decision feedback in conventional reduced-state decoders makes it impossible to directly use those high-speed design techniques.

We propose an approach, referred to as *quasi-reduced-state* trellis decoding, to eliminate the speed bottlenecks incurred by the decision feedback while largely maintaining the energy reduction potentials of reduced-state decoding. The basic idea is to only map the memory unit onto the reduced-state trellis while keeping the original full-state trellis for ACS computation. Again, let $N$ denote the number of states in the original trellis. Given $N = 2^{t_{MEM}} \cdot N^r_{MEM}$, where the positive integer $t_{MEM}$ is called the trellis quasi-reduction factor, we obtain an $N^r_{MEM}$-state trellis by grouping every $2^{t_{MEM}}$ states in the original trellis into a single state. The ACS computation and memory unit are mapped onto the original $N$-state trellis and the reduced-size $N^r_{MEM}$-state trellis, respectively, as illustrated in Fig. 2(a). A trellis matching block is used to accommodate the trellis size difference with the ratio of $2^t_{MEM}$ between the ACS computation and memory unit. As shown in Fig. 2(b), the trellis matching block contains $N^r_{MEM}$ data select blocks, where each one selects one out of the $2^t_{MEM}$ ACS outputs using a compare&select (CS) array.

Due to the absence of decision feedback, quasi-reduced-state decoders can achieve the same decoding throughput as a full-state decoder and directly enable the use of those high-speed design techniques for a significant speed improvement. From the decoding performance perspective, it is intuitive that its performance will sit in between that of full-state decoder and conventional reduced-state decoder.

## III. Importance-Aware Clock Skew Scheduling

The above proposed quasi-reduced-state decoding scheme eliminates the decision-feedback-induced speed bottleneck while maintaining the energy saving on the data storage block.

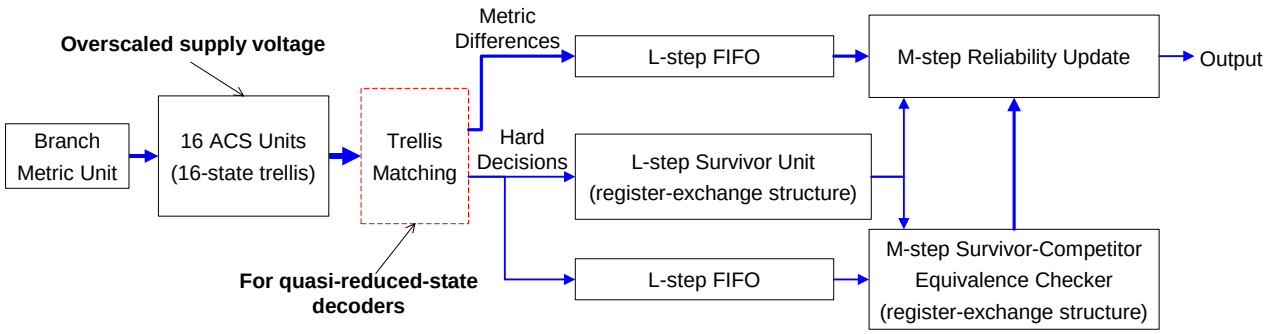


Fig. 3. System architecture of the SOVA decoders.

This technique can be directly integrated with an importance-aware clock skew scheduling technique that intends to reduce the trellis state metric computation energy consumption at small decoding performance degradation. The basic idea of importance-aware clock skew scheduling is briefly described as follows according to [8].

In a synchronous circuit, clock skew is defined as the time difference, $s_{i,j} = t_i - t_j$, between the clock arrival times $t_i$ and $t_j$ of two sequentially adjacent flip-flops (FFs), $FF_i$ and $FF_j$. Let $T_{CP}$ denote the clock period, and $D^{i,j}_{MAX}$ and $D^{i,j}_{min}$ denote the maximum and minimum propagation delays from $FF_i$ and $FF_j$, respectively. The value of clock skew $s_{i,j}$ should fall into $[-D^{i,j}_{min},\ T_{CP} - D^{i,j}_{MAX}]$ that is called permissible range [9]. To improve the circuit reliability, a safety margin presented by $M$ may be introduced between the clock skew and the ends of the permissible range. Clock skew scheduling refers to a process that optimizes the safety margins subject to certain criteria, which can be mathematically formulated and solved using various optimization techniques such as linear programming or some graphical methods. One of the most widely used clock skew formulations is shown as follows:

$$\begin{aligned} \text{Max} \quad & M \\ \text{Subject to}: \quad & s_{i,j} \le T_{CP} - D^{i,j}_{MAX} - M \\ & s_{i,j} \ge -D^{i,j}_{min} + M \end{aligned} \tag{1}$$

Based on the above conventional clock skew scheduling, we have an intuitive formulation for the importance-aware clock skew scheduling as follows:

$$\begin{aligned} \text{Max} \quad & M \\ \text{Subject to}: \quad & s_{i,j} \le T_{CP} - D^{i,j}_{MAX} - \gamma_j \cdot M \ , \\ & s_{i,j} \ge -D^{i,j}_{min} + \gamma_j \cdot M \end{aligned} \tag{2}$$

where the importance factor $\gamma_j \in (0, 1)$ quantitatively represents the importance of the destination signal of each individual path, and a more important signal has a larger importance factor. A semi-systematic method was presented in [8] for determining the importance factors.

## IV. Design Example

In this work, we use SOVA decoder for partial response channel as a test vehicle to evaluate the proposed quasi-reduced-state trellis decoding technique and its integration with importance-aware clock skew scheduling. We consider the enhanced extended partial response class 4 ($E^2$PR4) channel in the presence of additive white Gaussian noise (AWGN). The characteristic polynomial of $E^2$PR4 channel is $1 + 2D - 2D^3 - D^4$, which corresponds to a 16-state trellis structure. Hence, the full-state trellis decoder will map the SOVA algorithm onto this 16-state trellis, which is denoted as FS-S16 decoder. By setting the trellis quasi-reduction factor $t_{MEM}$ to be 1 and 2, we have two quasi-reduced state decoders denoted as QRS-S16-8 and QRS-S16-4, respectively.

Fig. 3 shows the system architecture of these decoders. The $L$-step survivor unit finds the survivor path using the register-exchange structure. The $M$-step survivor-competitor equivalence checker also has a register-exchange structure and exams whether the survivor and its competitor are equal or not at each trellis depth, based on which the reliability metric will be updated by the $M$-step reliability update block. For a detailed description of those functional blocks in Fig. 3, readers are referred to [7]. In our design, the parameters of $L$ and $M$ are set as 30 and 15, respectively.

To evaluate the SOVA decoding performance, we define the decoder output signal to noise ratio (SNR) as follows. Let $L_{ideal}$ denote the ideal soft output, i.e., the soft output of the full-state FS-S16 decoder without the presence of AWGN and voltage overscaling. For any decoder that operates in presence of AWGN (and voltage overscaling) and generates the soft output $L_{real}$, we define the corresponding output SNR as

$$10 \log_{10} \frac{\text{Mean}(L^2_{ideal})}{\text{Mean}(L^2_{real})}.$$

All the decoders are designed using Synopsys tool set with a 65nm CMOS standard cell library and operate at 1GHz clock frequency that leads to 1Gbps decoding throughput. Since the VOS-induced timing errors are highly dependent on the real circuit realizations, we carried out extensive post-synthesis circuit simulations to evaluate the decoder output SNR degradation when the supply voltage is overscaled. We use the Synopsys Composite Current Source (CCS) model [10] to enable the simulations under voltage overscaling. Because CCS is current-based, it can enable both temperature and voltage scaling of the cell behavior and achieve the timing analysis accuracy within 2% of SPICE [11].

When the importance-aware clock skew scheduling is applied to the decoders QRS-S16-8 and QRS-S16-4, we denote the corresponding decoders as QRS-S16-8-CS and QRS-S16-4-CS, respectively. Fig. 4 shows the decoder output SNR vs. voltage overscaling factor for QRS-S16-8-CS and QRS-S16-4-CS based on post-synthesis circuit simulations. For the purpose of comparison, Fig. 4 also shows the results when the importance-aware clock skew scheduling is not used for the two quasi-reduced-state decoders. It shows that the importance-aware clock skew scheduling enables a more graceful decoding performance degradation in presence of overscaled supply voltage.

Fig. 5 shows the simulated decoder output SNR vs. channel SNR for the decoders with different configurations. It shows that the proposed quasi-reduced-state decoding technique can very effectively reduce the overall decoder energy consumption at small performance degradation. When the importance-aware clock skew scheduling is further used, we may explore a wider spectrum of energy vs. performance design trade-offs.

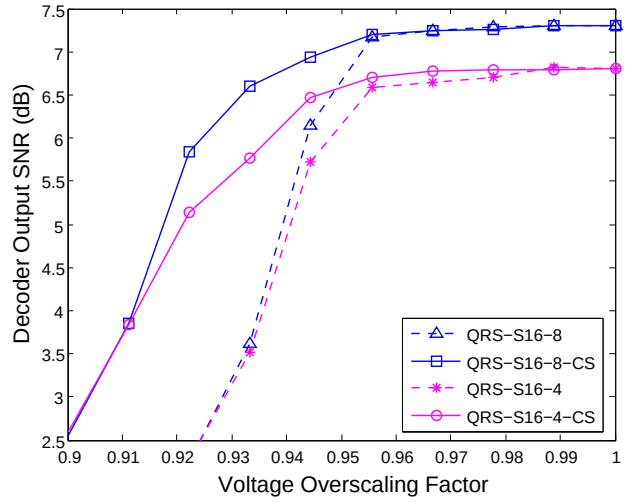


Fig. 4. Simulated decoder output SNR vs. supply voltage overscaling when channel SNR is 5.5dB.

## V. Conclusion

This paper presents a quasi-reduced-state trellis decoding technique to eliminate the decision-feedback-induced speed bottlenecks in conventional reduced-state trellis decoders, while largely maintaining the energy saving potentials. This algorithm-level technique can naturally integrate with a circuit-level importance-aware clock skew scheduling technique that enables the use of voltage overscaling on the state metric computation at small decoding performance degradation. These two techniques together will enable the designers to flexibly explore a wide spectrum of energy vs. performance trade-offs in soft-output trellis decoder design. This is further demonstrated using 1Gbps SOVA decoder design for partial response channel in magnetic recording channel as a test vehicle.

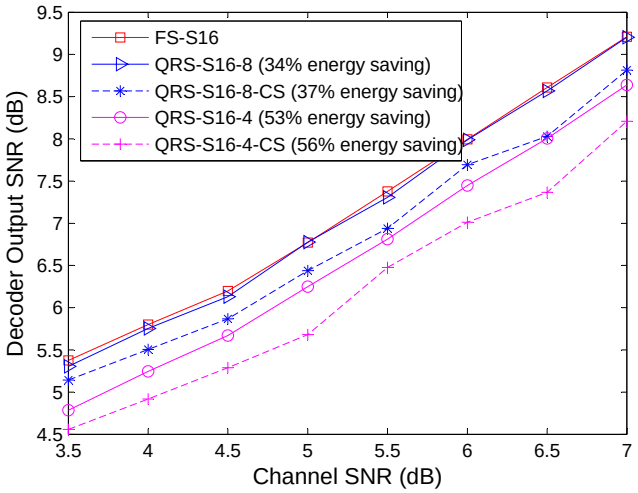


Fig. 5. Simulated decoder output SNR vs. channel SNR.

## References


[1] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 284–287, March 1974.

[2] J. Hagenauer and P. Hoeher, "A viterbi algorithm with soft-decision outputs and its applications," in *Proc. of IEEE Global Telecommunications Conference*, Nov. 1989, pp. 1680 – 1686.

[3] S.H. Muller, W.H. Gerstacker, and J.B. Huber, "Reduced-state soft-output trellis-equalization incorporating soft feedback," in *Proc. of Global Telecommunications Conference*, Nov. 1996, pp. 95–100.

[4] Z. Qin and K. C. Teh, "Iterative reduced-state decoding for coded partial-response channels," *IEEE Transactions on Magnetics*, vol. 41, pp. 4335–4337, Nov. 2005.

[5] G. Fettweis and H. Meyr, "High-speed parallel Viterbi decoding: algorithm and VLSI-architecture," *IEEE Communications Magazine*, vol. 29, pp. 46–55, May 1991.

[6] G. Fettweis, R. Karabed, P. H. Siegel, and H. K. Thapar, "Reduced-complexity Viterbi detector architectures for partial response signaling," in *Proc. of IEEE Global Telecommunications Conference*, Nov. 1995, p. 559563.

[7] E. Yeo et al., "A 500 Mb/s soft-output Viterbi decoder," *IEEE Journal on Solid-State Circuits*, vol. 38, pp. 1234–1241, July 2003.

[8] Y. Liu, T. Zhang, and J. Hu, "Low power trellis decoder with overscaled supply voltage," in *IEEE Workshop on Signal Processing Systems (SiPS): Design and Implementation*, 2006, pp. 205–208.

[9] J.L. Neves and E.G. Friedman, "Optimal clock skew scheduling tolerant to process variations," in *Proc. of Design Automation Conference (DAC)*, June 1996, pp. 623–628.

[10] SYNOPSYS Composite Current Source, http://www.synopsys.com/.

[11] George Mekhtarian, "Composite Current Source (CCS) Modeling Technology Backgrounder," Nov. 2005.