特集

ハイエンドプロセッサ内蔵SRAM技術

Embedded SRAM Technology for High-End Processors

あらまし

富士通は，社会基盤を支えるサーバ商品に搭載するプロセッサを国内では唯一独自開発している。その開発戦略は，半導体部門と協力してテクノロジと同時にプロセッサを並行開発するものである。ここで紹介するSRAM技術とは半導体製造と回路方式の複合技術であり，高性能，小面積，低消費電力の相反する要件を高いレベルで満足させ，テクノロジの立上げと同時に完全動作させるために必要な技術である。プロセッサのシステムクロック限界を律速するのは1次キャッシュ用高速SRAMの動作速度であり，処理のボトルネックとなる外部メモリのアクセス頻度を決めるのは2次キャッシュ用高密度SRAMの搭載容量である。この観点からSRAMはプロセッサのキーコンポーネントと言える。一方，半導体の微細化に伴い様々な弊害が顕在化し，SRAMの記憶素子であるメモリセルの製造ばらつきが増大している。このためサーバプロセッサの要件を満たせるようなSRAMの開発は非常に難易度が増している。本稿では，このような状況の中で，富士通がサーバプロセッサ用SRAMをどのように開発しているのかを紹介する。

Abstract

Fujitsu is the only company in Japan that develops its own processors for use in server products that support the social infrastructure. Its processor development strategy is to collaborate with the internal semiconductor group and simultaneously develop the processor and semiconductor technology. This paper introduces SRAM development technology, which is a complex technology combining both semiconductor manufacturing and circuit systems. It fully meets conflicting server processor requirements such as high performance, small area and low power. It is a technology that is essential for starting up new technology and having it fully operational at the same time. Level-1 cache SRAM speed determines the processor clock rate, while the data processing bottle-neck is determined by the density of the level-2 cache SRAM. Thus SRAM is a key technology for server processors. As finer semiconductor technologies progress, various problems arise and the variability of the memory cell in the SRAM gets bigger. Consequently, development of SRAM that meets the server processor requirements is getting critical. This paper describes our SRAM development methodology.

中台裕志（なかだい　ひろし）

エンタプライズサーバ事業本部テクノロジ開発統括部 所属
現在，サーバプロセッサ向けSRAMの開発に従事。

伊藤　学（いとう　がく）

エンタプライズサーバ事業本部テクノロジ開発統括部 所属
現在，サーバプロセッサ向け2次キャッシュ用高密度SRAMの開発に従事。

植竹俊行（うえたけ　としゆき）

エンタプライズサーバ事業本部テクノロジ開発統括部 所属
現在，サーバプロセッサ向け1次キャッシュ用高速SRAMの開発に従事。

ま え が き

富士通は，社会基盤を支えるサーバ商品に搭載するプロセッサを国内では唯一独自開発している。サーバプロセッサは高性能，高密度，低消費電力が求められており，その動作周波数限界を律速するのは1次キャッシュ用SRAMであり，処理のボトルネックとなる外部メモリのアクセス頻度を決めるのは2次キャッシュ用SRAMの搭載容量である。この観点から，SRAMはプロセッサのキーコンポーネントと言え，プロセッサと同様の相反する要件に対して高度に最適化を図ることが必要となる。SRAMは記憶素子であるメモリセルとこれを制御する周辺回路から成り，この構成を常に見直すことで最適化を実現してきた。

一方，半導体技術はムーアの法則に従って微細化を続けているが，SRAMのメモリセルはこの微細化に伴う不可避な製造ばらつきの増大によって，メモリセルとして求められる性能と安定性は低下の一途をたどっている。

このような状況の中でサーバプロセッサの要件を満たせるようなSRAMの開発は非常に困難になりつつある。本稿では，この技術課題の説明とこの解決に向けた富士通の取組みについて紹介する。

SRAM開発における技術課題

半導体メーカ各社は先端技術を駆使して微細化競争を続けており，プロセス技術のシンボルとも言えるSRAMのメモリセルの面積をテクノロジごとに半減させ続けている。しかしメモリセルが小さくなることで，これを構成するトランジスタ素子の製造に必要な不純物の拡散のゆらぎや，形状の不均一性などの物理的に不可避な現象によって，素子特性ばらつきが増大してくる。この結果としてSRAM設計が非常に困難なものとなってきている。

一般に，SRAMのメモリセルの性能と安定性はトレードオフの関係にあり，メモリセルを構成するトランジスタの閾値(いき)を小さくすると性能は向上するが，安定性は低下してしまう。半導体の微細化による閾値のばらつき増大によって，従来に比べて安定性の悪いメモリセルが出現する確率が高くなる。そこで歩留まりを確保するために，素子全体の閾値を安定寄りに，すなわち性能低下側に設定せざるを得なくなる。したがって，メモリセルを半導体技術のトレンドどおりに小さくしていくことは相対的にメモリセルの性能が低下することであり，高速に動作させたいサーバプロセッサとしての要件を満たすことが困難になってきたことを意味する。

SRAM開発の取組み

このような状況において，サーバプロセッサ向けに高性能，高密度，低消費電力なSRAMを開発するために著者らは大きく下記の3点の取組みを行っている。

(1) サーバプロセッサに最適なメモリセルの開発

メモリセルは各世代の最小サイズのものが最適なものとは言えず，プロセッサの要件とSRAMの回路方式に応じて面積と性能の最適化が必要である。著者らはテクノロジ開発の早期から半導体部門と連携し，協同でサーバプロセッサに最適なメモリセルの開発を行っている。

(2) 微細化問題を解決できる回路技術の開発

SRAMはメモリセルとこれを制御する周辺回路から構成されている。著者らはメモリセルの制御方法に関して，先述の技術課題の解決に向けて技術動向調査や独自の研究を行っており，SRAMの技術ロードマップを策定している。これに添って製品用マクロ設計と並行して次世代に向けた先行試作実験を繰り返し，新回路技術の立上げに取り組んでいる。

(3) シミュレーション技術の向上

素子ばらつきを考慮した統計的なメモリセルのワーストケースモデルを算出し，これをSRAM全体の回路シミュレーションに反映することで，実際の試作実験の回数を減らすと同時に設計段階での品質を確保している。この手法によって半導体テクノロジの立上げと同時にプロセッサの完全動作を可能にしている。

以下，これら三つの取組みについて詳細に述べる。

最適なメモリセルの開発

まず，半導体の微細化によって派生する問題をより詳しく述べる。SPARC64 Ⅷfxプロセッサ[1]に採用した45 nm世代の高密度用SRAMのメモリセルとその等価回路を**図-1**に示す。図に示すように1 μm^2にも満たない領域に6個のトランジスタが配置され

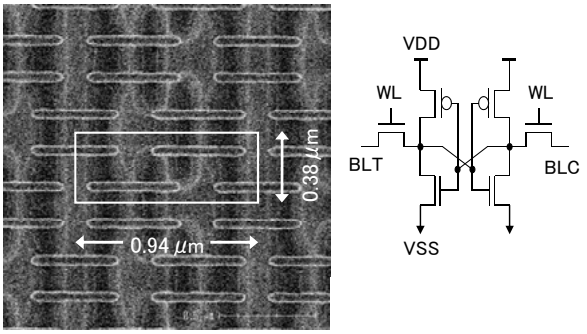
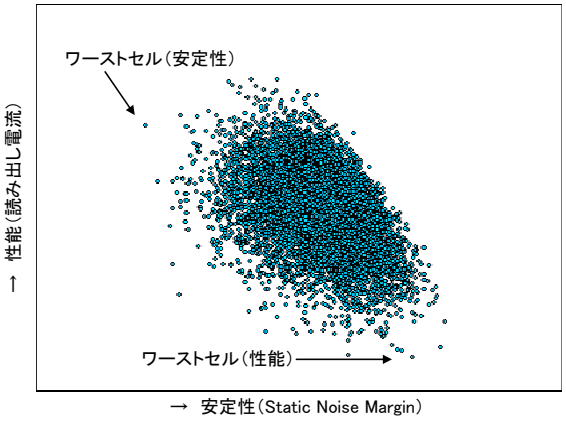
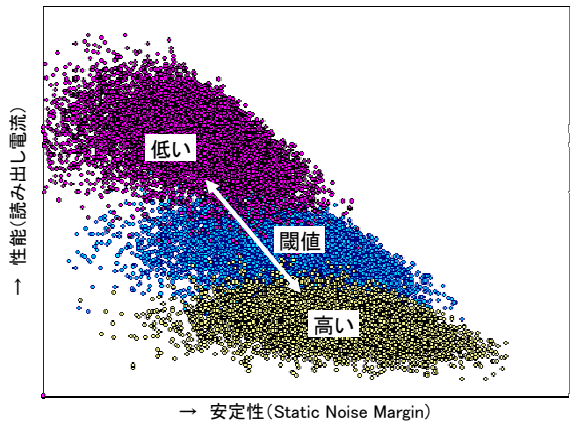


図-1　SRAMメモリセルのSEM画像と等価回路
Fig.1-SEM image of SRAM memory cell and equivalent circuit.



（a）メモリセルの性能vs安定性



（b）メモリセルの閾値調整による分布の変動

図-2　メモリセルの特性分布
Fig.2-Characteristic distribution of memory cell.

ている。2次キャッシュメモリはこれを数千万個並べて構成する。

このような微細なトランジスタを大量に使用することで製造ばらつきの影響を大きく受けるようになる。そこで，個々のトランジスタの製造ばらつきを反映させた場合にメモリセルの特性がどのように分布するかをシミュレーションした結果を**図-2**（a）に示す。縦軸はメモリセルの性能指標の一つである読み出し電流であり，横軸はメモリセルの安定性の指標であるSNM（Static Noise Margin）である。図中の点がメモリセル一つに対応する。このシミュレーションでは1万個のセルについてプロットしたが，プロセッサにはこの数千倍のメモリセルが含まれるため，実際にはより広範囲に分散する。プロセッサの性能は内蔵されるメモリセルの中に出現する最も悪いメモリセルが決定してしまい，安定性の最も悪いセルが歩留まりを決めてしまう。ばらつきが増大するということは図中の各点の分散が更に拡大することを意味する。

ばらつきの増大によって，最も安定性の悪いメモリセルのSNMが0以下になるとデータを保持できなくなる。これを回避するために素子の閾値を大きくすることでSNMを大きくすることが可能である。しかし，図-2（b）に示すとおり，性能を犠牲にしなければならなくなる。これが微細化に伴うメモリセルの問題である。

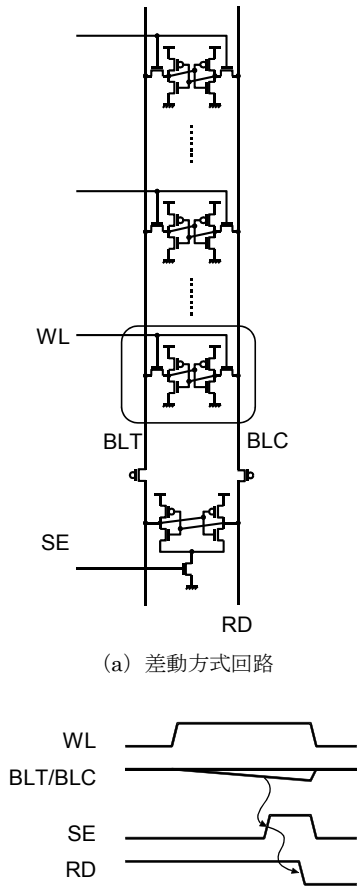
半導体メーカ各社は，先端技術を駆使して世代ごとにメモリセルの面積を半減させてきたが，そのテクノロジにおける最小のメモリセルはサーバプロセッサにとっては最適なものではない。SRAMはメモリセルとそれを制御する周辺回路から構成され，これらはお互いに密接に関係しており，SRAMに求められる性能，面積，電力の要件に応じて最適化が必要になる。例えば最小のメモリセルを使って高速動作を実現するためには，低下した読み出し電流によって減少した微小振幅でも読み出せるセンスアンプが必要になる。一般に入力振幅が小さくなるほど，これを増幅するセンスアンプの面積が大きくなり，結果としてメモリセルは小さいものの，SRAMとしては大きくなってしまう。さらに，大きなセンスアンプを駆動するため，消費電力も増大する。したがって，性能要件を満たしつつSRAMとしての面積が最小になるような読み出し電流を確保するために，あえて大きめなメモリセルを採用することが必要になる。このようにSRAMに求められる要件によって最適なメモリセルの面積は変わってくる。著者らは高速動作が必要な1次キャッシュ用SRAMと，面積優先の2次キャッシュ用SRAMそれぞれにおいて最適化を行い，結果として異なる面

積のメモリセルを採用した。さらにトランジスタの閾値を変えることで特性の最適化も図っている。

また，SRAMに求められる要件だけではなく，その回路構成によっても最適なメモリセルは異なってくる。そこで，著者らはテクノロジ開発の早期から半導体部門のメモリセル開発に参画している。

回路技術の開発

はじめに，従来から用いられている差動方式SRAMの読み出し動作について述べる。差動方式SRAMの回路図とタイミングチャートを**図-3**に示す。図中にBLT，BLCとして示されるBit線には多数のメモリセルが接続されている。メモリセルは二つのインバータの入出力を互いに結線することによりデータを保持している。ワード線WLを“1”にすることで読み出すべきメモリセルを選択し，保持データをBit線に伝播させる。しかしメモリセルは非常に小さなトランジスタで構成されており，多数のメモリセルが接続されているBit線を十分に駆動することができないため，BLT-BLC間の電位差が電源-GND間に比べ微小な振幅となる。そこで，この微小な振幅を増幅するためにイネーブル信号SEによりセンスアンプを活性化させ，出力信号RDに伝播させることで読み出し動作を完了する。



(a) 差動方式回路

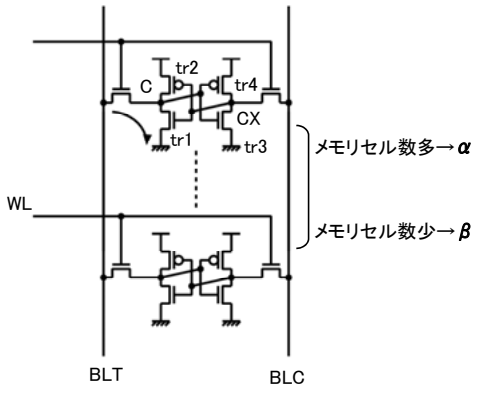
(b) 差動方式タイミングチャート

図-3　差動方式
Fig.3-Differential method.

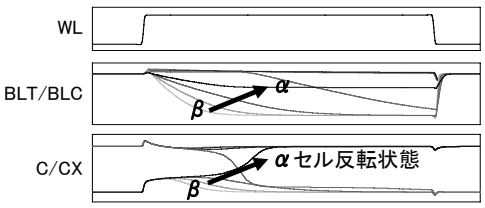
つぎに，著者らが導入した回路技術について述べる。微細化に伴うメモリセルの安定性低下の問題については学会などで様々な回避策が議論されている。著者らはメモリセルが駆動するBit線の負荷を軽く，すなわちBit線に接続するメモリセルの数を少なくし，読み出し時のBit線の放電を急速に行うことで，不安定に出来上がったメモリセルであっても反転を防止できることに着目した。これによって性能を犠牲にすることなくメモリセルを縮小できると考えた。メモリセルの回路図を**図-4**（a）に示す。また，製造ばらつきが大きく，反転しやすいメモリセルに対してBit線の負荷を変化させたときのシミュレーション波形を図-4（b）に示す。このようにBit線の負荷が大きい状態 α で読み出した場合，読み出し電流がBit線から流入することでノードCの電位は上昇する。これによってtr3とtr4で構成されるインバータが応答し反転することで，保持データが破壊される。

一方，Bit線の負荷を軽くした状態 β のような場合ではBit線の電位が急激に下がることで先述のインバータが反転する前に読み出しが完了し，データの反転を防ぐことができる。図-4（c）はBit線に接続するメモリセルの個数を変えることでBit線の負荷を変えた場合に，保持した値が反転する限界のばらつき量をシミュレーションで算出した結果である。この図はBit線に接続するメモリセル数を64としたときの反転限界のばらつき量 σ を1としてメモリセル数を変化させた場合の反転限界をプロットしたものである。この結果から明らかなように，Bit線に接続するメモリセルの数を減らすことで，より大きくばらついても反転しにくくなることが分かる。

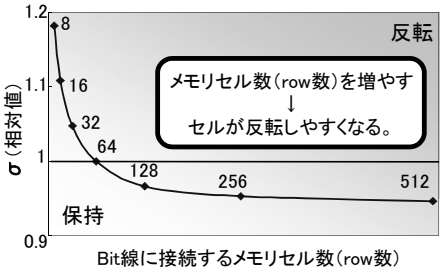
また，この効果を応用してSRAMの回路を考える場合，メモリセル数を十分少なくし，Bit線の振幅を大きくとることで，デジタル信号として扱うことが可能である。これによりSingle-End方式としてSRAMを構成することができる。



(a) 読み出し時のメモリセル動作

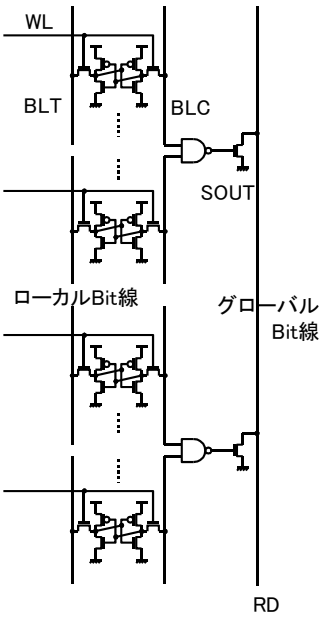


(b) 読み出し時のシミュレーション波形

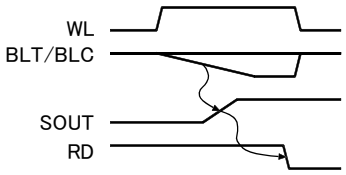


(c) Bit線負荷とセル安定性限界

図-4　メモリセルの安定性とセルアレイ構成の関係
Fig.4-Memory cell stability vs. cell array structure.



(a) Single-End方式回路



(b) Single-End方式タイミングチャート

図-5　Single-End方式
Fig.5-Single-End method.

ここで，**図-5**（a）と（b）にSingle-End方式の回路図とタイミングチャートを示す。Single-End方式では従来に対しBit線を分割するため負荷が軽くなり，読み出し時にワード線が開くとBit線はフルスイングする。このため差動センスアンプは不要となり通常のLogic Gateで読み出すことが可能である。この分割されたBit線をローカルBit線と呼ぶ。また，Bit線を分割したことによりデータを集約する必要が生じるが，これにはグローバルBit線を用いる。これらローカルとグローバルの2段階による読み出しにより，出力信号RDに出力を行う。

また，性能，面積，電力の面でも従来方式に比べて利点がある。

まず，性能面での利点を示す。Single-End方式ではBit線の長さを$1/N$とすることでメモリセルの放電時間を大幅に短縮することができるが，反面，ローカルBit線の制御回路がN個必要となり，データ集約のためのディレイのオーバヘッドが生じる。しかし従来のSRAMと比較した場合に，全体のディレイに対するセルの放電時間の比率が小さくなる。これによって，ばらつきが原因で遅いメモリセルが出現した場合の性能への影響を低減することが可能になる。

つぎに面積に関しては，従来の差動方式では微細化に伴うばらつき増大のために，センスアンプを小さくできなくなってきている。しかしSingle-End方式では制御回路の数がN倍になるものの通常のLogic Gateで構成することができるので，半導体の微細化に伴って通常の面積縮小率で小さくすることができる。

電力に関しては，読み出し/書き込みに必要な電荷（＝ダイナミック電流）が$1/N$となり，さらに

Bit線自体の負荷も1/Nになるのでこれを駆動するドライバも1/Nにでき，相乗的にダイナミック電流を減らすことが可能である。

シミュレーション技術の向上

メモリセルの性能は，SRAMの特性全体に対し，大きな影響を及ぼす。そのため，SRAMの設計時には，素子ばらつきを考慮し，実際に出現する最も特性の悪いメモリセルを精度良く算出しモデル化を行い，SRAMのシミュレーションに取り込む必要がある。

一般的に，ばらつきの推定手法として，モンテカルロ法が広く知られている。図-2に示したメモリセルの特性分布は，1万個のメモリセルの性能や安定性に関してモンテカルロ法によるシミュレーションで得られた結果である。しかし，実際のプロセッサには2次キャッシュメモリだけでも数千万個のメモリセルが搭載されており，この中に出現する最も特性の悪いメモリセル（ワーストセル）を，例えば1%以内の誤差率で精度良く算出するためには，数十億回以上のシミュレーションが必要になり，計算機資源や時間の制約によって事実上不可能である。

そこで富士通研究所が開発したSRAM解析システムを，ワーストセルの算出に応用することで，この問題を解決した。

この解析システムでは，まず，ばらつき係数を振り，マージンが減少する方向に進みながら，ワーストセルの探索を行う。つぎにISMC（Importance Sampling Monte Carlo）法(2)を用い，探索したセル近辺で集中的に乱数を発生させ重点的にサンプリングを行う。このとき，発生する多次元乱数には，各次元でのサンプリング点の配置が均等になるLatin Hypercube Sampling(3)を用いる。

これらにより，数百回程度のシミュレーション回数で，高精度なワーストセルの算出が可能となり，従来のモンテカルロ法と比較して百万倍以上の計算時間の短縮を実現した。

上記システムにより算出したワーストセルのモデルを取り込み，SRAMのシミュレーションを精度良く実行することで，試作回数を減らし，設計品質の向上を図っている。

む　す　び

本稿では，サーバプロセッサ向けSRAM開発の技術的課題と，この解決に向けた三つの取組みについて紹介した。

著者らは，これら取組みにより，45 nm世代において，SPARC64 Ⅷfxプロセッサの1次キャッシュメモリにSingle-End技術を適用した。さらに2次キャッシュメモリにも展開する予定である。

今後も継続して半導体微細化の課題解決に取り組み，高速，高密度，低消費電力なSRAMを開発することで，サーバプロセッサの性能向上に貢献していく。

参 考 文 献

(1) T. Maruyama : SPARC64™ Ⅷfx : Fujitsu's New Generation Octo Core Processor for PETA Scale Computing. Hot Chips 21，2009.

(2) R. Kanj et al. : Mixture importance sampling and its application to the analysis of SRAM designs in the presence of rare failure events. DAC 2006, p.69-72.

(3) A. Olsson et al. : On Latin hypercube sampling for structural reliability analysis. *Structural Safety*, Vol.25，Issue 1，p.47-68（2003）.