COMBINING SWITCH AND SITE YIELDS
FOR SOFT-CONFIGURABLE WSI

Miriam Blatt
Sun Microsystems Inc. M/S 17-10
2530 Garcia Ave., Mountain View 94043
mblatt@vis.eng.sun.com

ABSTRACT
Soft-Configurable wafer yield models depend on both switch and site yields. Earlier articles presented analytic models showing yield loss due to switch and site failures. Monte Carlo simulation combines these components by setting switches and sites faulty according to their expected yields, and attempting configuration. Redundant columns provide backup for either switch or site failures as needed. Simulation results show that both spare columns and switch bypasses significantly reduce the dependence of wafer yield on switch yield.

INTRODUCTION
A vital part of every Wafer Scale Integration (WSI) design is its method of circumventing manufacturing defects. Without fault tolerance no working wafers can be produced. With fault tolerance, some working wafers are produced, but their number (called yield or harvest) depends on the application requirements. Requiring few faults produces low wafer yields, while tolerating many faults produces high wafer yield. Such high yield results from accepting a system with fewer working components, thus lowering wafer area utilization.

Soft-Configurable WSI (as introduced in [1]) provides a network of switches to communicate amongst the fault-free application sites. Soft-configurable switches reduce fabrication costs by requiring only standard equipment, while allowing reconfiguration at runtime. Similar switch networks were designed for the ELSA project [2], and the Anamartic wafer memories [3].

Earlier articles [4,5] presented a prototype pipelined memory system demonstrating our approach to Soft-Configurable WSI, analysing wafer yield in two phases. Monte Carlo simulation based on fabrication defect statistics predicts the yield of switch and site components, using the VLASIC yield simulator [6]. Models were developed to predict wafer yield loss due to either switch or site component failures, however these analytic models could not produce a combined wafer yield that includes the effects of both switch and site failures.

We describe the configuration requirements of a soft-configurable pipelined memory. Wafer yield models are introduced, with a description of the Monte Carlo combined simulation. Graphs of simulator results show that both spare columns and switch bypasses significantly reduce the dependence of wafer yield on switch yield.
Figure 1: Block diagram showing connections between memory sites (rectangles) and soft-configurable switches (circles). Arrows show potential links that are enabled only between fault-free units.

**CONFIGURATION OF A WSI PIPELINED MEMORY**

Figure 1 shows the block diagram of a pipelined memory prototype implemented to demonstrate the Soft-Configurable switch network. Sites (shown as rectangles) hold the memory blocks, and switches (shown as circles) control the routing of data by setting control lines for the interfaces between them (shown as arrows). Addresses, data, and configuration commands enter from switches at the bottom, and data exits through switches at the top. Configuration commands establish a vertical switch path, that is tested independently of the memory sites. Faulty switches are bypassed using horizontal links between the good switches. After switch path configuration, memory sites adjacent to the paths are tested.

Switch path configuration sets mask bits telling each switch which neighbors to trust. Path configuration is protected by a rotation code that detects stuck and bridging faults. Sites also use mask bits to guard against faulty switches. This design is described in detail by Blatt [7].

Throughput is increased by simultaneous use of multiple paths in addition to pipelining within each path. Built on a larger scale, it could be used as a main memory providing interleaving (using multiple paths) and fast fetches of cache lines (by pipelining within a path). Memory was chosen as an example vehicle that is readily implemented. Network switching may be even better suited to this approach, as described by Katevenis [8].

**YIELD EVALUATION MODELS**

Switch and site circuit yields provide inputs to various models for evaluating wafer yield. Circuit yields were obtained through Monte Carlo simulation based on the prototype layouts and defect statistics. Results reported earlier [4,5] have been modified to include recently obtained data on defect size distributions [9]. This data (from wire test pattern yields) shows defect rates for larger defects falling considerably slower than the default continuous distribution used by the VLASIC simulator that assumes a $1/z^3$ curve. The modified defect size distribution averages three measured fabrication laboratory results; it is shown together with the updated data in Figure 2. All inputs other than the size distribution remain as
Defect Tolerance

Circuit I Transistors I Yield(%) switch 1266 99.2
4K RAM 28K 93.3
16K RAM 104K 78.8
64K RAM 403K 40.9

Figure 2: Defect Size Distribution and Circuit Yield Results

described in [4,5]. The switch yield dropped from 99.5\% to 99.2\%, and the 4K RAM yield dropped from 96.4\% to 93.3\%. The 16K and 64K results are extrapolated assuming a layout similar to the 4K RAM used in the prototype.

Wafer yield curves presented in [4] and [5] rely on analytic models that derive wafer yields from the probability that a given number of functioning components can be found in a useful arrangement. For instance, the balanced load model requires a minimum number of good sites per switch path, assuming that switch paths can be configured to access all sites within a pair of site columns.

Switch paths are restricted to the three switch columns surrounding a pair of site columns. They can split and rejoin to bypass faulty switches, setting mask bits to ignore outputs from faulty neighbor switches or sites.

Yield curves presented in [4] assume very high switch yield, as switch effects are brushed aside with the promise of over 99\% simulated yield. The assumption that switch yield can be disregarded is examined more closely in [5], showing the gains from one or two extra columns, and from allowing limited bypassing for faulty switches. Results presented in [5] show that wafer yield due to switch failures drops rapidly as switch yield decreases, requiring extremely high switch yields unless extra columns and/or bypasses are permitted. Extra columns increase the silicon area overhead, and bypasses increase the delay along the switch path.

Combining the analytic models of [4] and [5] to produce wafer yields based on both site and switch yields is complicated by interdependencies — switch failures may leave sites inaccessible, and site failures may leave a working switch path without enough functioning sites. Monte Carlo simulation handles these interdependencies by randomly assigning switches and sites to be faulty according to their respective yield values, and then checking for switch paths with the minimum required number of working sites. This produces wafer yields for the balanced load model described above, with spare columns providing backup for either switch or site failures as needed.

Each data point combines the results of 5000 iterations, followed by groups of 500 iterations until the yield changes by less than 0.0008 (absolute). Results presented in 1989 by Harden [10] show close agreement between Monte Carlo simulation and analytical methods using 2000 to 3000 iterations.
YIELD RESULTS

Wafer yield is presented as a function of site yield together with either switch yield or the number of spare columns. Figure 3 uses straight switch paths only, while Figure 4 allows arbitrary bypasses within the 3 switch columns surrounding each pair of site columns. Straight vertical paths minimize latency, showing the limiting case where horizontal links are ignored.

The top two graphs of Figure 3 show wafer yields when the number of spare columns varies, and switch paths are straight. Clearly, 98% switch yield severely limits yield for the larger wafers with 16 rows. There is a significant improvement with each additional spare column. With 8 rows, additional spare columns beyond 2 add little improvement.

The bottom two graphs of Figure 3 show wafer yields when the switch yield varies, and switch paths are straight. These curves show how sensitive wafer yield is to small changes in switch yield. This sensitivity is much more extreme for the wafers with 16 rows. Switch yield must exceed 99% for wafer yield to reach acceptable values, without providing more than 2 spare columns.

The top two graphs of Figure 4 show wafer yields when the number of spare columns varies, allowing switch path bypasses. These curves are not limited by the 98% switch yield; 2 spare columns provide wafer yields over 99.7% once site yield exceeds 90%.

The bottom two graphs of Figure 4 show wafer yields when the switch yield varies, allowing switch path bypasses. Notice that switch yield varies down to 89.5%, rather than 95.5% as in Figure 3. With bypasses, switch yield is much less critical, providing wafer yields over 98% for switch yields over 95.5% and site yields over 90%. This provides a reasonable margin for switch design complexity.

In every case, wafer yield rises to a plateau. The plateau level is limited if there are too few spare columns, or if switch yield is low. The larger wafer with 16 rows has a longer plateau followed by a steeper drop. Compared to site data without switch yields, these curves have a lower maximum, dependent on the switch yield, and fall faster as switch yield loss eliminates access to functioning sites.

Switch path bypasses add one or more extra clock phases of delay. Odd numbers result from bypasses at the input or output port (but not both). Initial estimates of the number of switch paths requiring multiple bypasses indicate that the vast majority need no more than two or sometimes three extra phases delay. When neighboring switch paths bypass using the same switches, switches are shared between the two paths. This was not checked, since external software can ensure that data is not sent at the same time to two switch paths when a shared piece has a conflict.

In [5], bypasses were assumed to be two sided, and to bypass only one switch within the path (not a port). These pessimistic assumptions produced lower wafer yields. The Monte Carlo simulation checks for every possible bypass, including bypasses of multiple faulty switches, and half bypasses where only one side has enough good switches, but there are enough good sites accessible from that side. Port bypasses are permitted, but required to be half bypasses, even if both nearby switches are working. This assumes an external interface with a predetermined number of switch paths per wafer.

Bypass optimization selects the switch path that minimizes additional delay. Coalescing multiple bypasses reduces the number of horizontal hops. Single as well as double bypasses
Figure 3: Varying the number of spare columns and the switch yield, with straight switch paths. $P_{SW}$ is the switch yield, $f$ is the fraction of sites required per column-pair, $r$ and $c$ are the number of rows and columns, and $(c - r)$ is the number of spare columns.
Figure 4: Varying the number of spare columns and the switch yield, with switch path bypasses. $P_{sw}$ is the switch yield, $f$ is the fraction of sites required per column-pair, $r$ and $c$ are the number of rows and columns, and $(c - r)$ is the number of spare columns.
can be coalesced in those cases where the join is short enough so that there are enough good sites accessible. When both switches near a faulty port are functioning, the half bypass can be chosen to minimize delay. Optimized paths combined with Monte Carlo simulation could show wafer yield as a function of maximum bypass delay, where the maximum extra delay is set to 1, 2 or 3 clock phases.

CONCLUSION

Monte Carlo simulations show that wafer yield dependence on high switch yields can be reduced substantially using switch path bypasses and/or spare columns. These simulations combine the effects of switch and site failures in ways that cannot easily be captured using simpler analytic models, due to interdependencies between the effects of switch and site failures on system configuration.

REFERENCES