Backend Flow Optimization
Using Design Structure Matrix

Badari Kommandur
Prithvi Bannerjee

Courtesy of Badari Kommandur and Prithvi Bannerjee. Used with permission.
Overview

• Introduction
• Motivation for backend flow optimization
• CPU Design Phases
• Silicon Debug Process Flow
• Design Structure Matrix Application to Backend Process
• Managing Unplanned Iterations
• Key Learnings
• Recommendations
Motivation

- Design philosophy aims at reducing risk with time
- Significant source of uncertainty
  - Market Needs
  - Technology (Design & Manufacturing)
  - Resource roll offs
- Capital commitment needs to be made before uncertainty resolves
- Backup options hard to pursue in parallel

Significant capital has to be committed based on demand forecasts and technology investment needs to be made well before uncertainty has resolved.

Modeling and prototypes not sufficient to model real designs sufficiently.

Objective is to reduce risks as design moves into development and manufacturing rapidly. At any time instant, at least 3 generations of CPU designs are in different stages of execution. (i.e There may be 3 designs each for IA-32, IA-64 and mobile platforms). These teams are geographically distributed and may be located in US (multiple sites), Malaysia, India, Israel. A given design may be developed concurrently across sites.
Motivation

- Backend flow optimization is critical for revenue growth
  - Reduce TT$
  - Performance Enhancement
  - Yield Improvement
  - Reduce Lead time for Demand (MCT Reduction)
- Intel does not use formal methods for IPT formation
  - Use of DSM to validate current composition of IPT
- Analyze process to reduce time spent in backend flow
  - Reduce iterations?

Assuming a annual microprocessor shipment rate of 100 million units, ASP (Average selling price) of $200, Yield Impact to revenue is calculated. Added opportunity cost and capacity to meet demand can result in total impact to be ~2X

Tapeout to PTQ (Production tape quality) ➔ First time units can be sold to customers for revenue
Methodology

- Interviews with technical and management leads in Pentium 4™ design team (Design, Manufacturing, Planning, QRE)
- Integration of a priori knowledge about backend flow
  - Hands on experience on previous generations of CPU designs
- Review of methodology/process documentation for Product Development Teams
- Analysis of functional effort spent on last 4 generation of IA-32 processors (Lead designs and Compaction designs)
- Design Structure Analysis of Backend Flow
Front end development involves the integration of University research, Marketing, OEM, Lead Users, Finance, HR, Executive Staff, Strategic Planning, Platform Solutions Group, Test Technology and Manufacturing groups in defining the lead architecture.

Product roadmap planning involves continuous input from the competitive intelligence group, OEMs and strategic planning group. These can result frequently in roadmap changes to react to market environment and competitive position. (Pentium 4™ HT Extreme Edition launch before Opteron™ launch)

Product Implementation Plan results in generation of External Product Specification (released to customers) and Internal Product Specification based on which design execution starts.

Backend development involves tight integration of all functional teams through Integrated Product Teams, War Rooms and Taskforce meetings. Typically takes 3-4 quarters from tapeout of lead architecture to HVM. Quality and Reliability engineering works closely with design teams and PDT to certify design for launch while meeting DPM goals.
There are two types of microprocessor projects at Intel. Lead Architecture Projects and Compaction. Lead Architecture involve radical innovations in design, process and manufacturing. Initial manufacturing is on N-1 process generation for short duration of time before migration to lead process technology. These are used to validate design concepts and process technology.

Compaction involve minor modifications to architecture and major focus is on incremental innovations to improve performance, yield and reduce cost for high volume manufacturing. These are manufactured on the latest generation of process technology.

Average throughput time for a lead architecture is about 4-5 years and constant effort is ongoing to reduce it. The CPU development effort can be broken into four main phases.

• Technology Readiness
• Front End Development
• Execution
• Backend (Silicon Debug/Ramp)

Lead B does not show first two phases but still is called Lead Design due to significant architectural and design changes incorporated based on product roadmap changes.

Lead Design takes approximately twice the design effort as a compaction project.
Die Preparation

Wafer Mount

Laser Scribe

Saw and Wash

Tape and Reel Die Sort

Wafer Mount: To keep the die from scattering when the wafer is cut

Laser Scribe: To create trenches in the wafer metal moats prior to Saw. This eliminates ILD layer delamination/cracking which can occur when the trenches are not created

Saw and Wash: To separate the individual dice on the wafer by cutting, and to wash away the cutting debris.

Tape and Reel Die Sort: To remove the good dice from the sawn wafer and place them in Tape & Reels, and bin them accordingly
Auto Package Load: To place package substrates into stainless steel carriers for processing

Flux Reflow: To form the electrical and mechanical connections between the die and package by reflowing the solder to form a high-quality solder joint

Epoxy Underfill Dispense: To fill the cavity under the die and around the die perimeter with epoxy. This seals the area and provides mechanical support for the die-to-package interconnects

Epoxy Cure: To complete the fillet process, and thermally set the epoxy underfill between the die and substrate
IHS Cure: To thermally set the sealant and IHS Thermal Interface Material to provide a strong thermal and mechanical connection between the die and the heat spreader

IHS Visual Inspection: To inspect the packages to verify the IHS quality

Laser Mark: To scribe 2D matrix in the identifying information on the package. This step follows shortly after IHS because the IHS process covers the substrate mark

Burn In: Dies are tested at elevated temperatures and voltage to screen for infant mortality. Tests are run at low frequency (aim to ensure toggle coverage)

HIS=Integrated Heat Spreader
Class (Structural and Functional Test): To perform 100% electrical test after Burn-In in order to isolate manufacturing defects, ensure product meets performance specifications per product data sheet, categorize components according to device performance (bins), and provide yield analysis/improvement feedback for Fab & Assembly.

Fusing: To set the speed of the microprocessor by blowing a fuse in the circuitry that controls the device clock rate. This step is done here rather than at Test to postpone the configuration of the CPU to the latest point in the supply line. This gives Intel the most flexibility in setting the CPU performance based on a specific customer order.

PPV: To verify that the microprocessors can run the various operating platforms they are designed to run (such as Windows, Unix, and OS/2). This is used to catch “test holes” (functional problems not caught at ST/FT Test).
Silicon debug process involves close coordination between the design engineers, product engineers and the fab engineers. Time is of the essence and any design escape can be very expensive at this stage.

More than 1 year of preparation/planning is required before the first silicon comes through the fab for test/debug operations.

There are multiple iteration loops between first silicon and test/debug operations.

PE: Product engineer
DE: Design engineer
Backend flow includes the iterative process of manufacturing and testing of a microprocessor. This is the key iterative process that is followed for performance improvement, yield enhancement and testing of a design.
There are three main iteration loops as shown in the DSM:

1. Sort loop
2. Quality Validation Loop (Class/Quality Loop)
3. PTQ Validation Loop

For the lead design there are planned iterations (called dash steppings) which aim to resolve the findings from the Si debug activities through the use of metal only changes (use of higher layer metal above m1 is preferred due to inline material in the fab).

Sort data is obtained at the wafer level (with x,y coordinates of each die) at cold temperature since heat dissipation capability is limited since there is no IHS or heat sink attach possible. Sort data is used to plot a Fmax v/s Isb curve. Extrapolation based on past sort to class correlation is used to predict the bin split which is a key input for design execution to improve the design to achieve higher performance through planned iterations (1 every quarter).

Class data is obtained on packaged parts at full speed and at hot temperature to emulated system level conditions as per spec. Customer Failure Analysis Correlation Requests can also generate design changes and addition of new test vectors to screen for marginal dies. The frequency of such iterations is very remote and usually require a “All hands on deck” approach to resolve it in the shortest time possible.

Class data results in identification of two kinds of limiters:

1. Speedpaths – Design works at lower speed than spec
2. Functional Failures- Design fails to operate at any frequency

Since class is done at elevated temperature, speedpaths and functional failures sensitive to leakage increase (exponential increase in temperature), coupling and memory cell instability issues are flagged.
DSM for Backend Flow

- Backend Flow represents the bulk of the iterative loops in the overall CPU development process
- Sequencing of tasks already optimum
- Three critical loops identified –
  - Sort loop
  - Quality Validation Loop
  - PTQ Validation Loop
PDT Product Development Teams

PDT is the central coordinating body which has member participation from all functional groups. All decisions that affect engineering commitments must be discussed in Co-ordination Meeting or War Room.

Escalation Path: In the event of serious disagreement (e.g. project priority), the escalation path starts with Product Coordination Meeting, War Room or Program Ops Meeting, and may be forwarded to RDM if appropriate.

Stepping WG focus on stepping specific issues. Stepping usually involve minor design changes to improve performance or functionality through either metal or all layer changes. (A1, A2, A3….Metal Only Changes B1, B2, B3….refer to all layer changes and make WIP useless)

Design Process Sync is weekly IPT whose primary focus is on design + process optimization to meet product roadmap (long range strategic planning/technical discussion forum)
Managing Iterations

- Speedpath Czars
- FMEA is done in parallel to disseminate learning to all projects outside of critical path
- Integration of on-die, design for test/manufacturing features critical to reduce impact of iterations
  - Clock tuning, Metal options, Bonus devices, Software controlled timing, DFT features...
- Live Methodology Documents
  - Updated through Techforum review
- Design Forums
  - Disseminate learning from backend flow to front end design
Key Learnings

- Intel has an inherited culture of relying on data driven, IPT based approach in product development
  - IPT membership has required membership to manage interactions effectively
- Sequencing of tasks in the backend is optimized based on past generation.
- Unplanned Iterations are managed by
  - Improved coordination through IPT
  - DFM, DFT features
  - Knowledge Sharing
- DSM adoption for new CPU projects can be critical to speed up NPI

New Product Introduction (NPI)
DFM = Design for manufacturing features
DFT = Design for test features
Pre Compensated Clock Skew
Due to variation in channel length, threshold voltage, supply gradient and other device parameters, clock skew is introduced in the design.
Compensation mechanism exist to correct this to reduce iterations post manufacturing as part of the test flow. The same mechanism is used to optimize the bin split at the highest frequency.
As show here, the max skew of 30 ps is reduced to less than 10 ps through clock tree de-skewing. Note that the 10ps offset is intentional skew to maximize bin split.

Post Compensated Clock Skew
Zero-skew clock not always best for max-frequency on silicon
Dominant post-silicon paths may be isolated to a few clock domains or phase dependant
Genetic algorithm developed on tester to run thousands of combinations skewing clock domains and duty cycle for best combination
Recommendations

• Product launch plans can be pulled in through improvements in Sort capability
  – Use of Hot Sort capability can result in early bin split information (~2 weeks lead time reduction in getting Fmax/Isb data at hot temp)
  – Enables changes to design earlier for new product launch sooner
• High Frequency of IPT meetings (daily) may impact normal work progress to some extent

DFT = Design for test features like probe node area control diodes to enable laser voltage probing for silicon debug. Critical for functional and timing debug
IPT = Integrated Product Teams

In the past due to daily frequency of War Rooms and high level of visibility in these forums, there have been occasions wherein normal work processes were neglected just to sustain progress updates in the IPT forums. This can be really expensive in stalling work progress through normal work channels.

Use of Hot Sort capability enhancements enable us to get hot temperature Fmax/Isb data to predict functionality and speed/yield information at least 2 weeks early (This emulates class results which are obtained after packaged parts are available for testing). This enables early progress in making design changes to fix functional or speedpath bugs. This can also assist in addition of tests to screen for marginal dies which fail at higher temperatures after class testing. This saves additional time since we can get a more robust test program for class ready before packaged parts arrive for testing.

Total effort saving is hence much more than 2 weeks (can be 2X of lead time savings)
Cd = Critical dimension or channel length
Based on Cd targets, the Standby current varies. Isb increases exponentially with Cd reduction
Shorter Cd result in faster devices and higher frequency of operation
Questions?
Backup
Commitment of resources is managed from phase to phase
Slanted lines between phases represent overlap of activities
Each phase includes inputs, management approval and documented results
Market and business needs drive the creation of product line families…
Test and Silicon debug not included in this rollup. Design engineering and mask design is the major resource sink ~70% of effort

Validation accounts for 1/3 rd of effort
ECO: Engineering change order. Contains detailed bug information and proposed fix. Options to verify fix with metal options or software workaround listed.

ECO Verification is done simultaneously through metal options under class testing and using simulation tools in the design database.

Learnings from each speedpath analysis are documented in a speedpath template and disseminated through design forums. Design rule checks are updated to verify entire database. Cross project wide dissemination of learnings is also enabled.
Overlap of Backend Steps

- T/O Mask
- 50 wfr P1 Out of Sort
- Dash T/O/Mask
- 50 wfr P1 Out of Sort
- ATD Assy & PDL Test
- ATD Assy & Validation
- Marketing QS
- Production Wafer Starts
- MTL 1
- PG A/T
- Production CDD

System Project Management – Fall 2003
DSM maps the entire product development flow.

Individual DSM developed to track the Accuracy, Completeness, Timeliness and Stability of the interactions.

Flows for information, material and resources are tracked for each interaction.