From Planning to Operations: The Ever-Shrinking Optimization Time Horizon

Deriving Benefit from Increased Solver Power

- Revisit previously shelved applications
- Build bigger, more accurate models
  - *Example:* Recent supply-chain model with 10 million constraints, 19 million variables (solve in 1.5 hours)
- Optimize "globally", over entities that were previously treated separately
- ***Move from the traditional Operations Research domain of planning to (real-time) operations: Business execution***

Planning versus Execution

- Planning (traditional OR application)
  - Data is typically aggregated
    - Accuracy issues can often be finessed
  - Decision cycles months or years
  - Emphasis of what-if analysis and decision “support”
- Execution
  - Data must be accurate
  - Decision cycles can be seconds to minutes
  - Solutions computed by software are often implemented as is

Planning versus Execution

- Planning
  - Pros
    - Easier to explain, control, use (run by experts)
  - Cons
    - Business impact is often obscured
    - Hard to maintain
- Execution
  - Cons
    - Harder to explain and control (not run by experts)
  - Pros
    - Direct business impact can be significant
    - System maintenance – you have no choice

Three Success Stories

Tales from the cutting edge

Ann Bixby, Brian Downs & Mike Self, Interfaces, Vol. 36, No. 1, January-February 2006, pp 69-86

The dance of the thirty-ton trucks

Karla Hoffman & Martin Durbin, *Operations Research*, Vol. 56, No. 1, January-February 2008, pp. 3-19

Short-interval production-line scheduling for front-end semiconductor Fabs

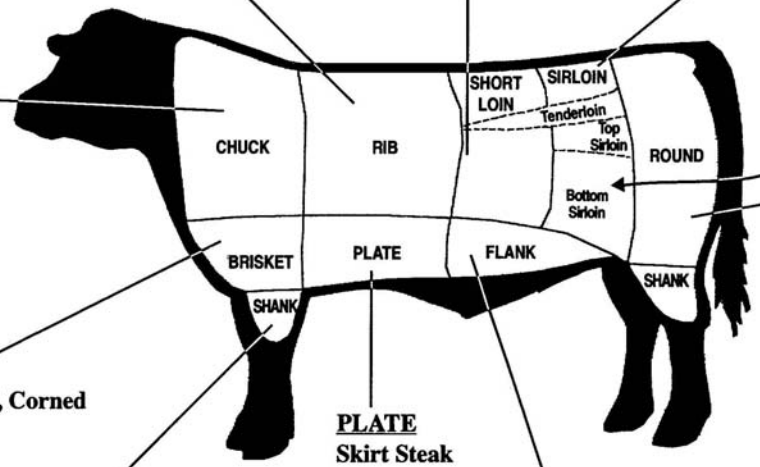
Robert Bixby, Rich Burda, Dave Miller & Steve Roberts, *Proceedings of ASMC* 2006, pp. 148-154

Remarks

- Each of these applications uses optimization
  - Linear and Mixed-Integer Programming
- I verified that all of these applications really are being used.
- Question: Did increased solving power really made a difference? Could we have done this 5-10 years ago?

Tales from the Cutting Edge:
A Scheduling and Capable-to-Promise Application for Swift & Company

— BEEF CUTS —
Where They Come From



CHUCK
Chuck Eye Roast, Boneless
Top Blade Steak, Boneless
Arm Pot Roast
Shoulder Pot Roast, Boneless
Mock Tender Roast
Blade Roast
Under Blade Pot Roast
7-Bone Pot Roast
Short Ribs
Flanken-Style Ribs
Cross Rib Pot Roast

RIB
Rib Roast, Large End
Rib Roast, Small End
Rib Steak, Small End
Rib Eye Steak
Rib Eye Roast
Back Ribs

SHORT LOIN *
Top Loin Steak, Boneless
T-Bone Steak
Porterhouse Steak
Tenderloin Roast/Steak

SIRLOIN *
Top Sirloin Steak
Sirloin Steak
Tenderloin Roast/Steak
Beef Tri-Tip
The tri-tip roast is a boneless cut from the bottom sirloin. It is also called a "triangle" roast because of its shape.

ROUND *
Round Steak
Top Round Roast
Top Round Steak
Bottom Round Roast
Tip Roast Cap Off
Eye Round Roast
Tip Steak
Boneless Rump Roast

BRISKET
Whole Brisket
Brisket, Point Half, Corned
Brisket, Flat Half

PLATE
Skirt Steak

FLANK *
Flank Steak
Flank Steak Rolls

SHANK
Shank Cross Cut

BEEF

National Cattlemen's Beef Association
444 North Michigan Avenue
Chicago, Illinois 60611
(312) 467-5520

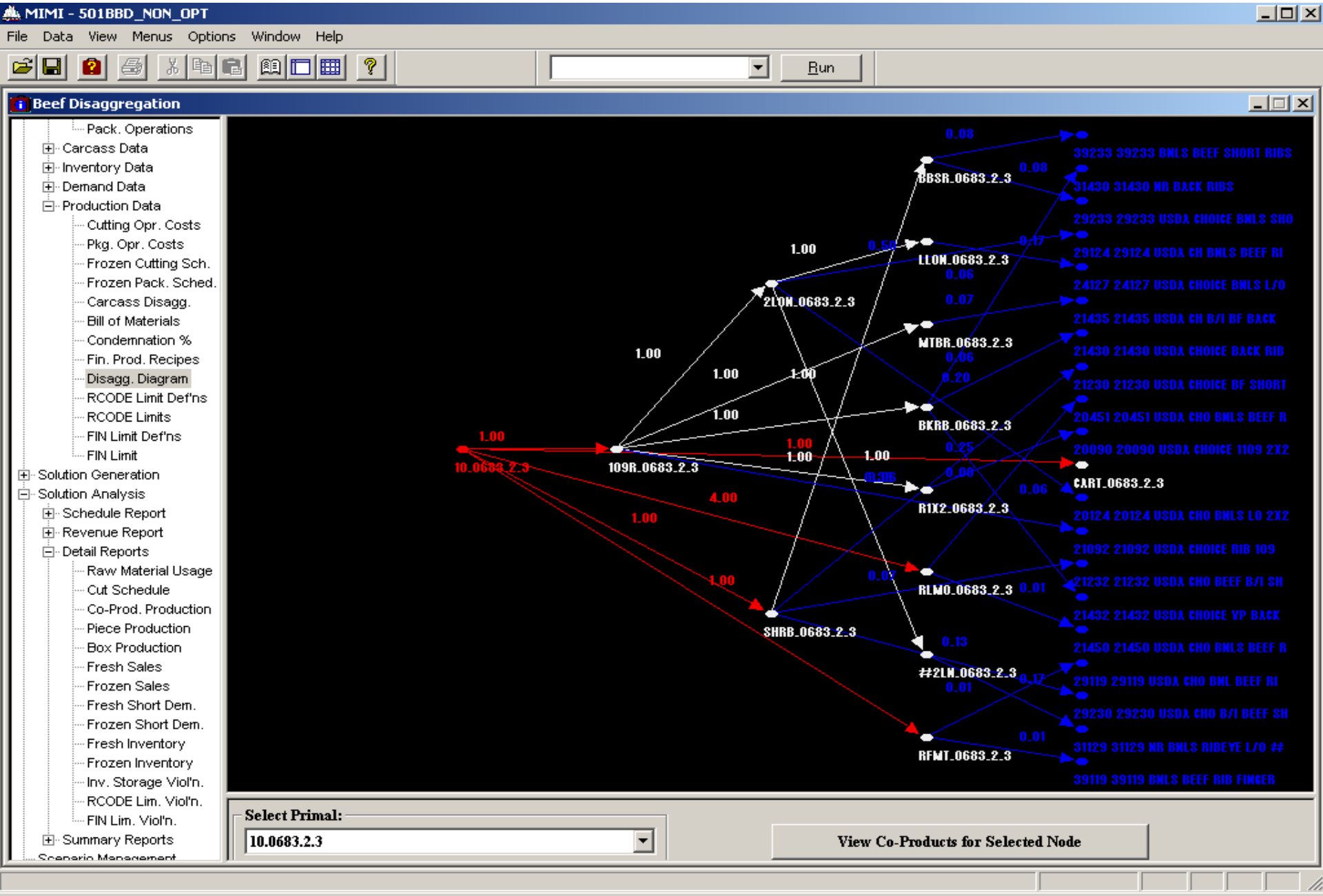
* Beef primals that feature cuts lowest in fat.

21-500

Beef Disaggregation

- The problem
  - 5 meat processing plants
  - Carcass inventory at each plant at shift start must be processed by shift end. Cut into 7 primals, USDA graded, "disaggregated" into pieces, and packaged.
  - This process must be scheduled, taking into account existing orders and *current forecast.*
  - Schedule must interact with the sales process.

A Carcass Disaggregation Tree



What Drove the Application

- The process
  - The schedule decides for each carcass a full disaggregation and packaging plan.
  - When you take an order, you would like to know what you are "capable" of supplying, not just what's in the schedule. This requires "moving up the tree": HUMANS can't do it – not during a sales call!
- The result
  - Lost sales, unfulfilled orders, dissatisfied customers.

Beef Disaggregation

- Started as 1 million variable “textbook” LP model.
  - After one year of model reductions (many very complex), the model was reduced to meet memory and *resolve-time limits (< 10 seconds)*
- The Environment:
  - 300 queries and commits (LPs) handled per hour by each model
  - A total of 45 models are running fully automated handling queries and commits 24 hours per day
- The savings:
  - $13 million/year (determined by internal benefits study)
  - Inventory sold increased from 10% to 80%
  - Most important: Business changed fundamentally

A Model Instance – LP

- Resolve-time requirement: <10 seconds
  - Model sizes: 250,000 constraints and 300,000 variables
- Query solve: Resolve from advanced basis with a small number of added rows and columns
  - CPLEX 9.0 (2004) 0.7 secs
  - CPLEX 5.0 (1997) 1.2 secs
  - CPLEX 1.0 (1988) 4.4 secs
- Machine speed adjustment:
  - CPLEX 5.0 (1997 PC -- 20x slower) 24 secs

Was increased solving power essential to this application?

The Dance of the 30-Ton Trucks:
Real-time dispatching of concrete trucks for Virginia Concrete

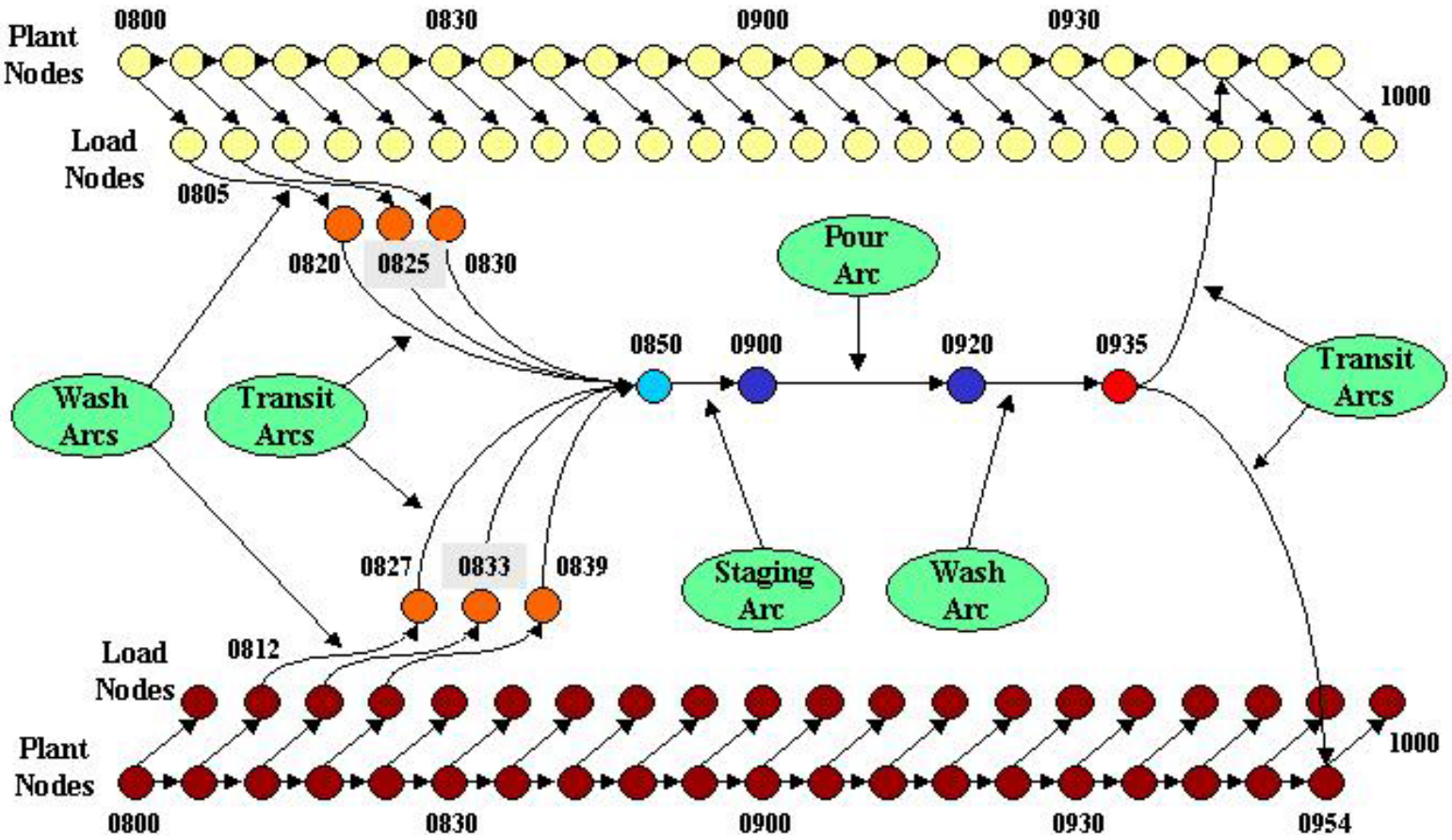
Concrete Delivery

- The Background
  - Virginia Concrete is a part of Florida Rock.
  - They deliver 500-700 loads per day to 150 customers, a total of 5000-6000 cubic yards of concrete per day.
  - Deliveries occur from 10 plants with 125-150 trucks.
- A key characteristic of the business
  - 90% of orders change before being delivered ➔ The delivery schedule is always out-of-date.
- The key driver for this application
  - The recognition that GPS provided a potentially very valuable technology for their business.
  - The result: A major program to introduce GPS technology and the necessary IT infrastructure.

The Optimization Solution

- Developing the solution
  - The initial expectation: Based upon experience, heuristics were expected to be the only viable approach.
  - The plan: Being aware of the advances in LP/MIP technology, at least give it a try.

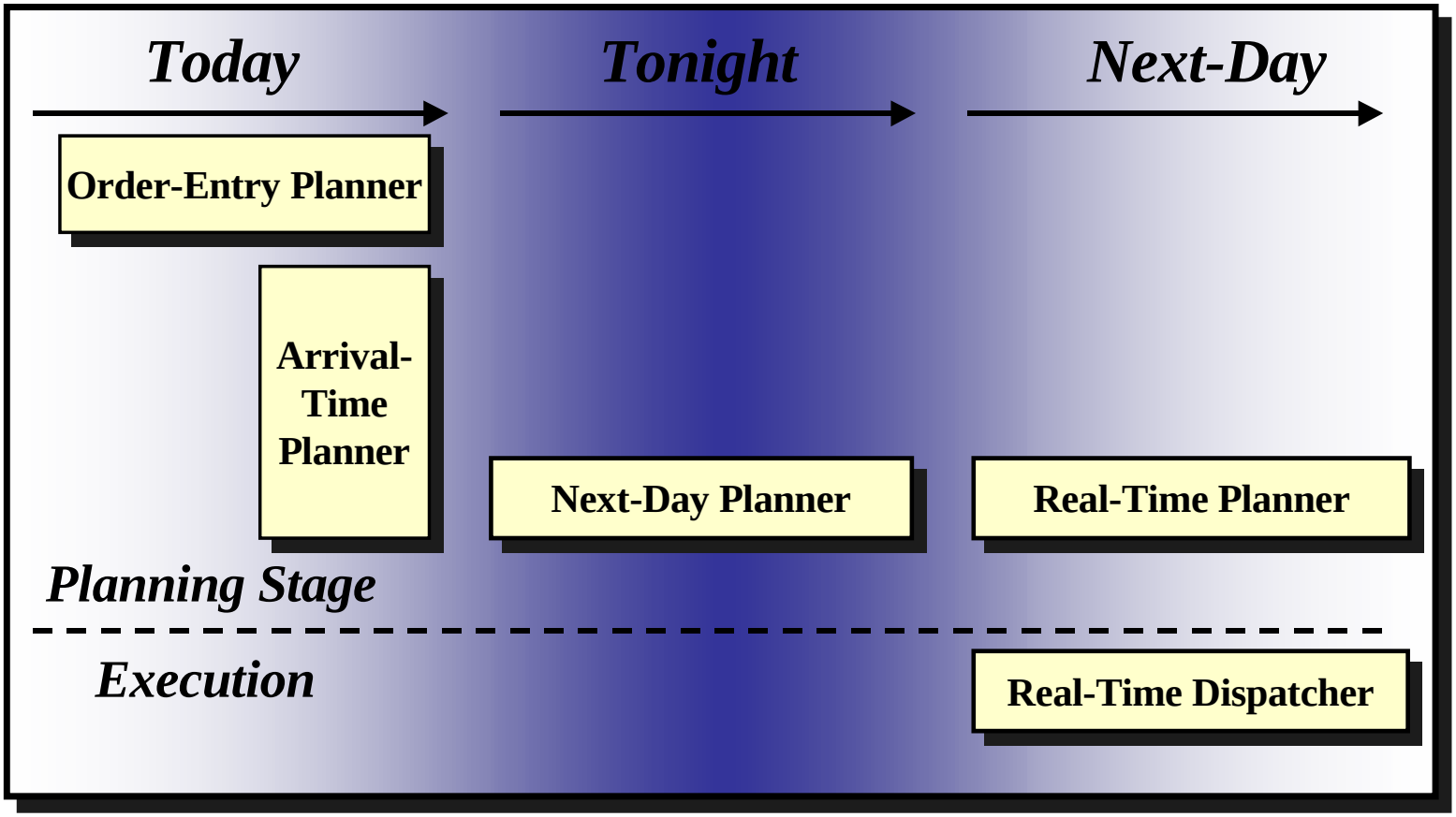
The Model Structure:
A space-time network



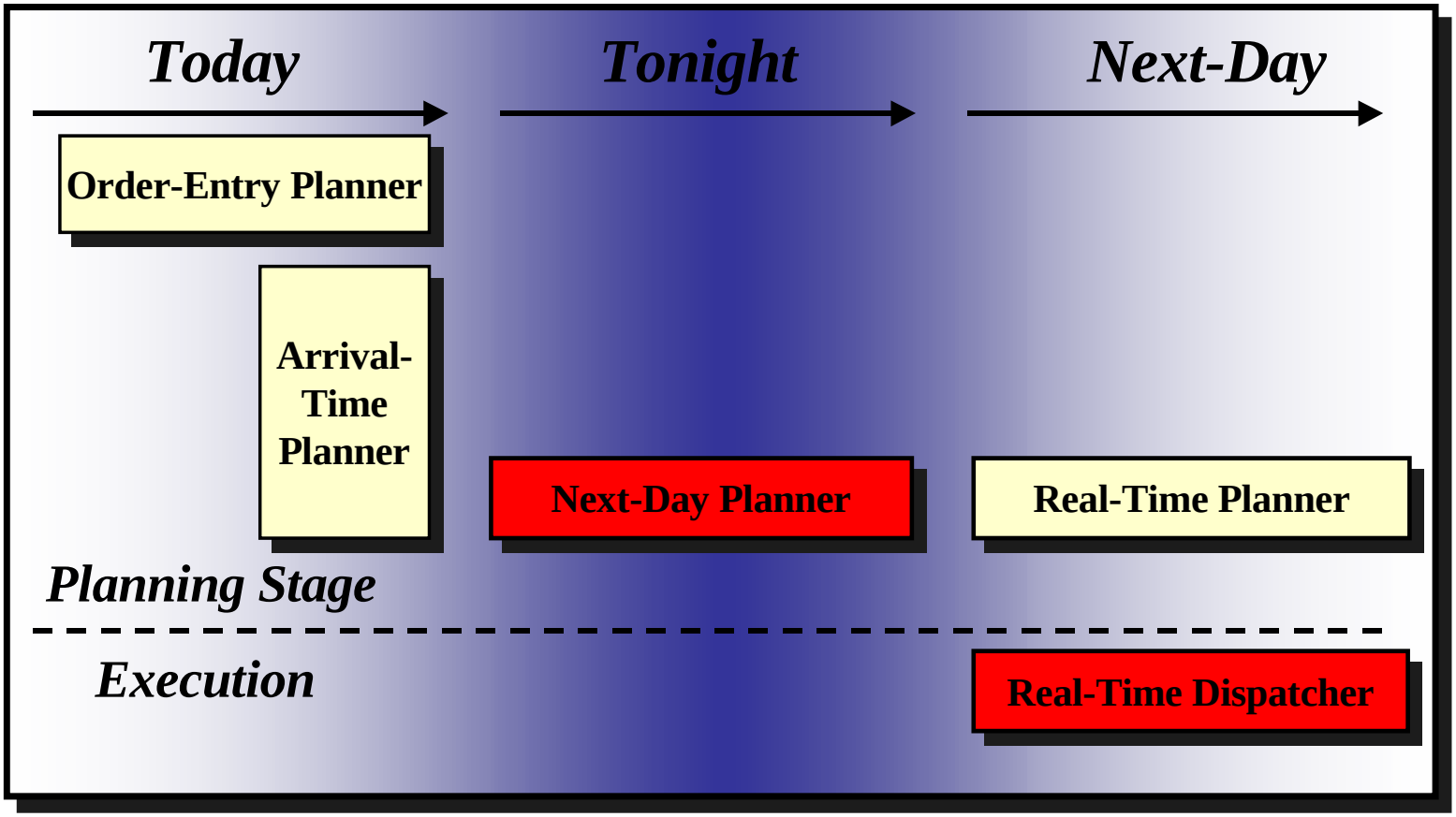
The Optimization Solution

- The characteristics of the solution
  - The key business benefits:
    - Employee retention through reduced stress on dispatchers.
    - A fundamental change from Truck-Based to Demand-Based dispatching.
  - The key OR modeling contribution: Dealing with infeasibilities.

The Decision Support Tool



The Decision Support Tool



Key Modules

Return on Investment

- Benefits
  - Eliminated employee retention problems
  - Quality of schedule less dependent on dispatchers
  - Schedule is now DEMAND-based rather than TRUCK-based (estimated savings of $750,000/year)
- Florida Rock is expanding and promoting this application
  - Now being deployed company wide (10x increase in trucks and plants)
  - FR is promoting industry wide as scheduling best practice

Model Instances

- **Model sizes:**
  - Next-Day Planner: 25000 cons, 200000 vars (2000 binary)
    Time Window to solve = 2 hours (4 hours accepted)
  - Real-Time Dispatcher: 10000 cons, 75000 vars (300 binary)
    Time Window to solve = 15 seconds (30 seconds accepted)

- **Summary: Where LP/MIP technology progress made a difference:**
  A. Dual simplex algorithm
  B. Heuristics in MIP

- **Next-Day Planner LPs – Solving the Root:**
  CPLEX 1.0 (1988) primal >40 hrs
  CPLEX 3.0 (1994) dual 18 mins
  CPLEX 9.0 (2004) dual 12 mins

Model Instances

Next-Day Planner MIPs – 2 hour window

Algorithm	Mean Time	First Solution
CPLEX 5.0 (1997)	5.1 hrs	4.1 hrs
CPLEX 8.1 (2003)	0.8 hrs	0.2 hrs

Real-Time Dispatching MIPs – 15 second window

Time Limit	15 secs	30 secs	60 secs
CPLEX 5.0	no feasibles	20% feasible	80% feasible
CPLEX 8.1 gaps	10.3%	1.5%	0.05%

Short-Interval Detailed Production Scheduling in 300mm Semiconductor Manufacturing

Robert Bixby

Other contributors to this work:
Vincent Gosselin (ILOG)
Rich Burda (IBM), Dave Miller (IBM)
Ed Rothberg (Gurobi Optimization)

Overview

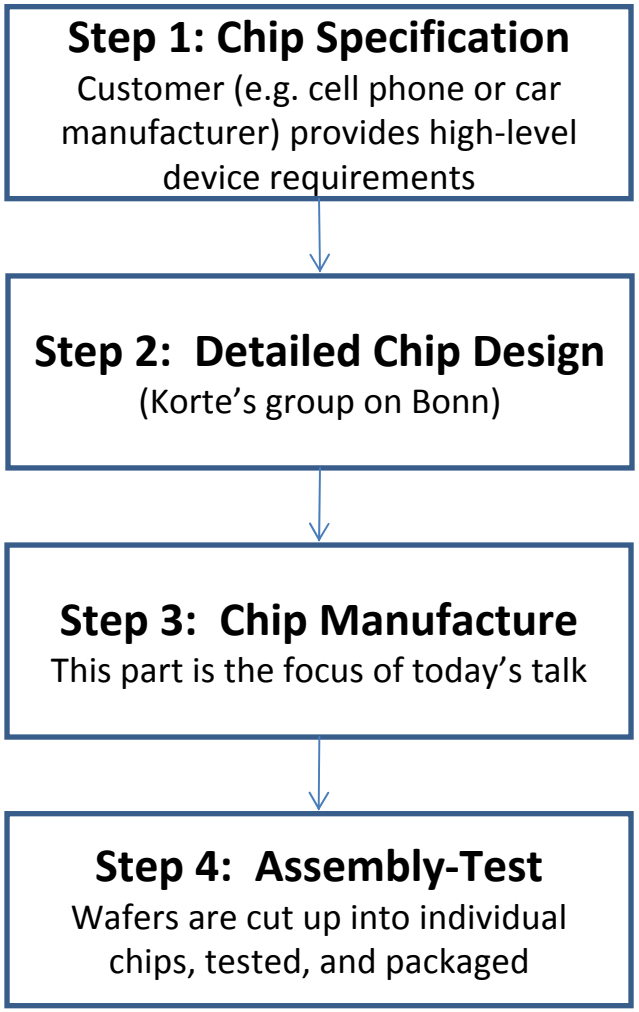
- Semiconductor manufacturing – background
- The scheduling problem
- ILOG Fab scheduling solution
- Benefits resulting from implementing ILOG solution

Semiconductor History

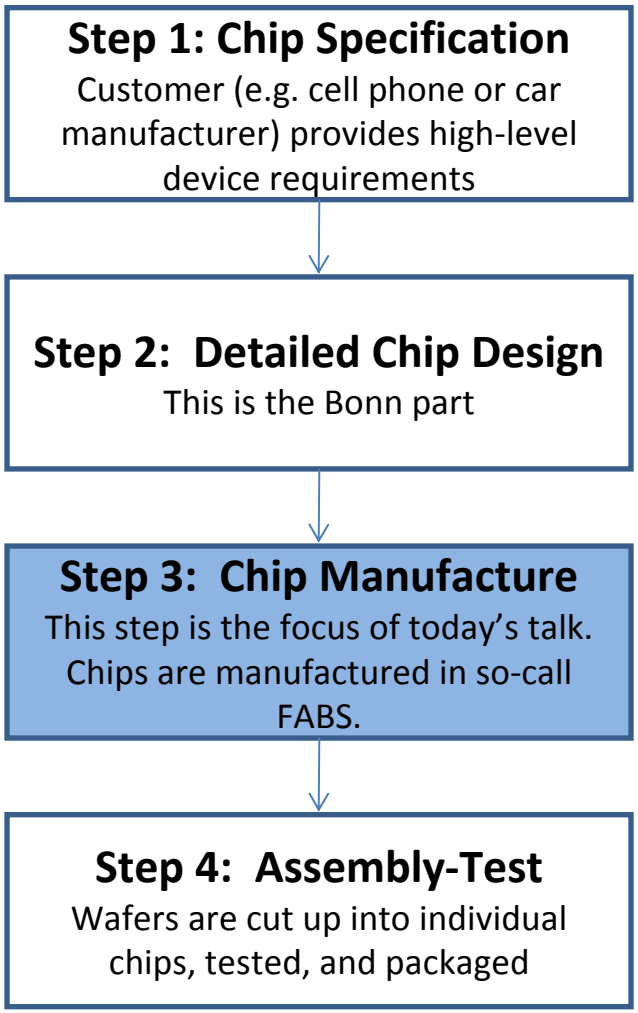
- 1947 Transistor invented
  - Bardeen, Brattin, Shockley at Bell Labs
- 1958 Integrated circuit introduced – circuits on a single, planar substrate
  - Kilby (TI), Noyce (Fairchild)
- 1960s – 90s Manufacturing processes revolutionized
  - 1964: Gordon Moore (Fairchild) predicted device density would double every 18 months
  - Rapid price drops began in mid sixties
- 1990 – Present: Focus on production issues
  - Automation
  - Cost control
  - Process control and efficiency

Semiconductor Manufacturing

The Semiconductor "Supply Chain"



The Semiconductor "Supply Chain"



Key Fab Performance Metrics

A brief Tutorial

Little's Law

Throughput = WIP / Cycle-Time

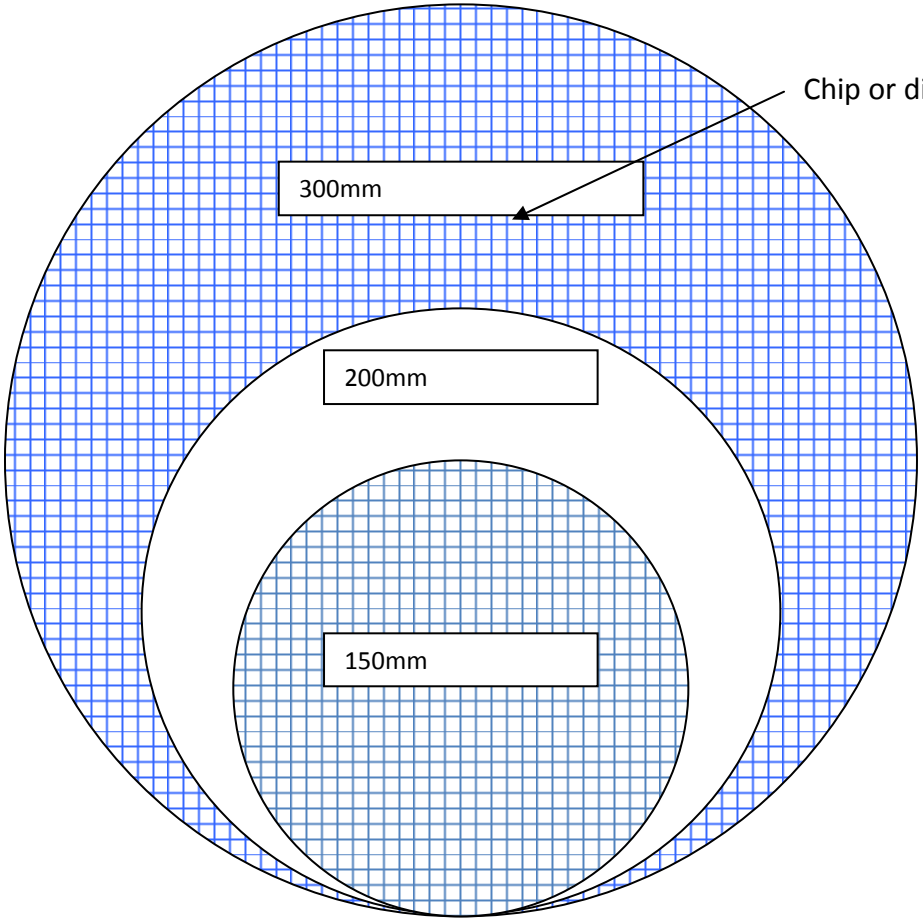
WIP = Work in Progress

Cycle Time = Wait time + Actual processing time

= Total processing time

The Holy Grail: Reducing Cycle Time

Silicon Wafers

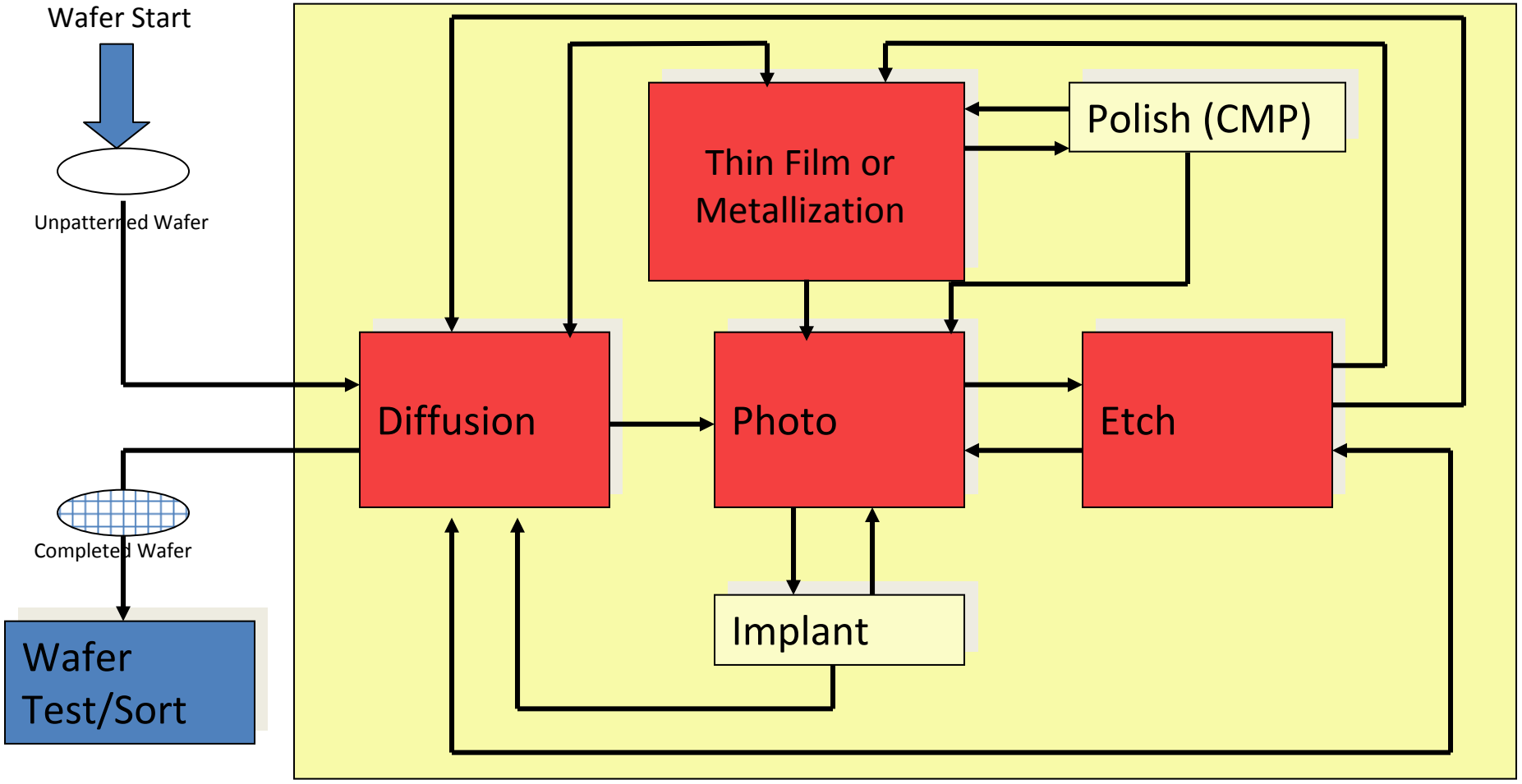


Some facts:

- **300 mm wafers current state-of-the-art**
- **500+ chips (dies) per wafer**
- **Process may require over 500 steps in 50 or more “layers”**
- **Wafers are processed in lots of 1-25 wafers**
- **Takes 1-3 months to process a lot**

A Re-entrant Fabrication Process

Main fab processes



The Scheduling Problem

Building 323 – IBM's 300 mm Fab

East Fishkill, New York

- Opened Summer 2002
- Cost $4-$5 billion
- **Fully automated** production environment
- **All lots are dispatched to tools without human intervention**

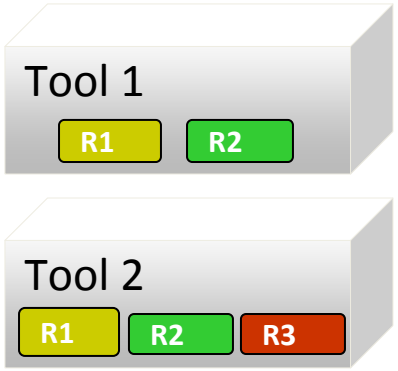
  15,000 dispatches per day

Current Industry Dispatching Solution
(Real Time)

- "Rule Based" – Heuristics
- "Opportunistic Scavenging"
  - Step 1: Tool announces that it needs work
  - Step 2: Dispatching system looks at queue of immediately available lots
  - Step 3:
    - Lots sorted by priorities, due dates, …
    - Rules of thumb applied to select from the sorted list
    - Real time checking – is the dispatch feasible?
    - Lot is dispatched

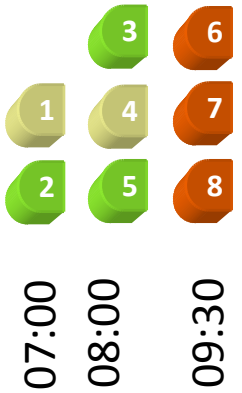
An Example

Tools & Recipes



Raw process time = 2 hours / lot

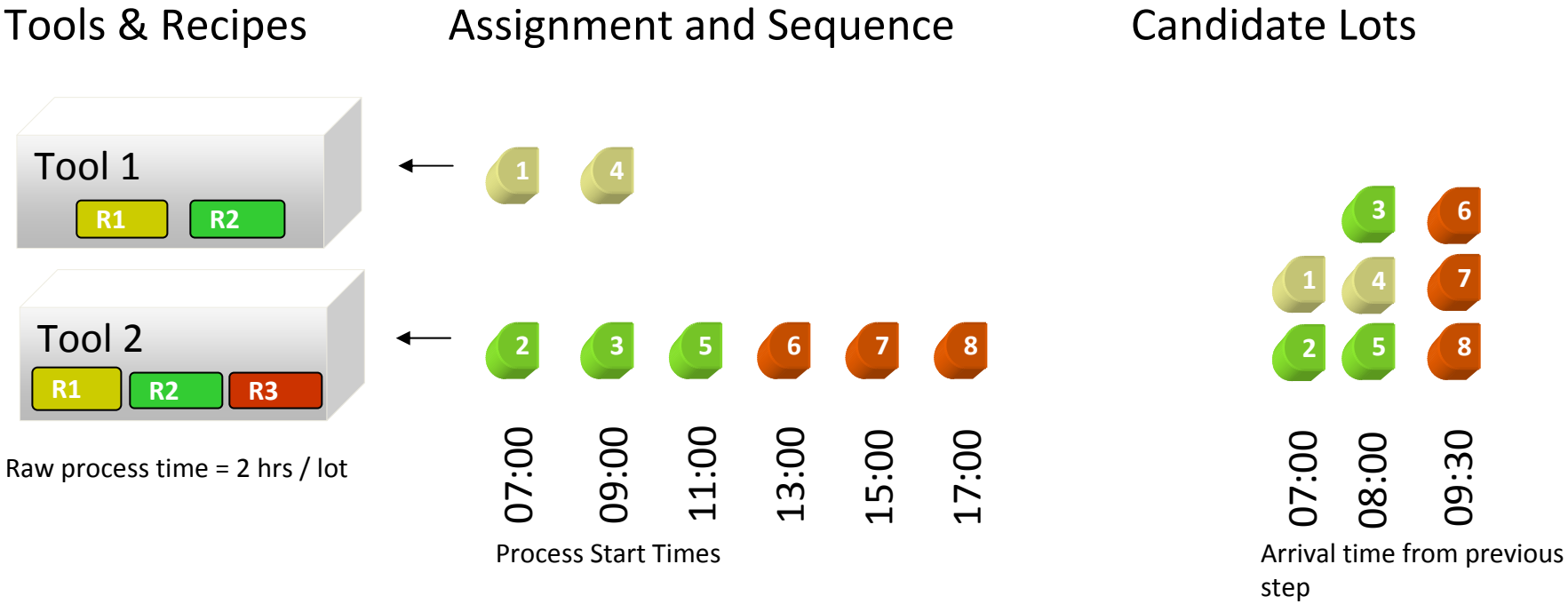
Candidate Lots



Arrival times
from previous step

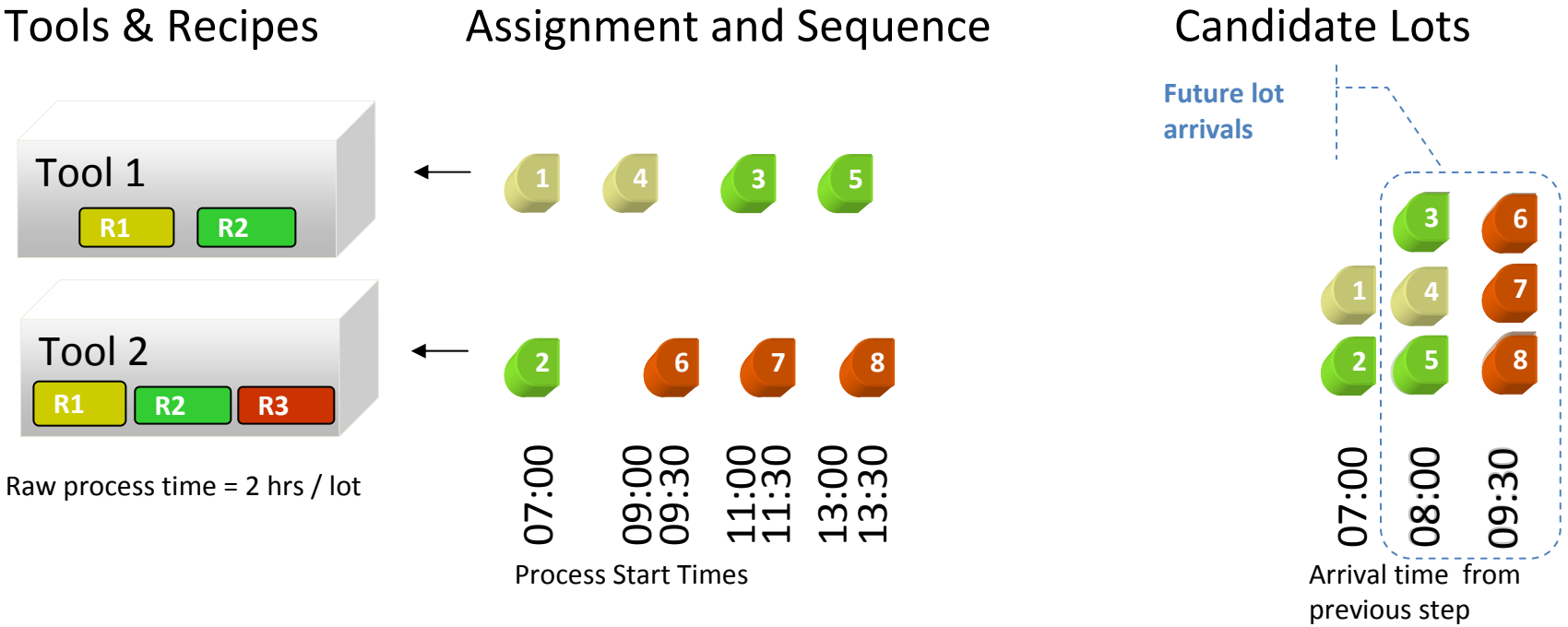
- For each process step, which tool should process each lot?
- For each tool, in what sequence should the lots be processed?

An Example



- Lot #8 cycle time = 9.5 hours
- Tool 1 utilization = 4/12 = 25%

An Example



Raw process time = 2 hrs / lot

- Lot #8 cycle time = 6 hours (37% improvement)
- Tool 1 utilization = 8/8.5 = 94% (73% improvement)

Advantages of Scheduling

- Advantages of scheduling vs. rules-based dispatch are well understood
  - Rules cannot see across tools
  - Rules have limited upstream vision
  - *Optimization automatically adjusts to changing business conditions*

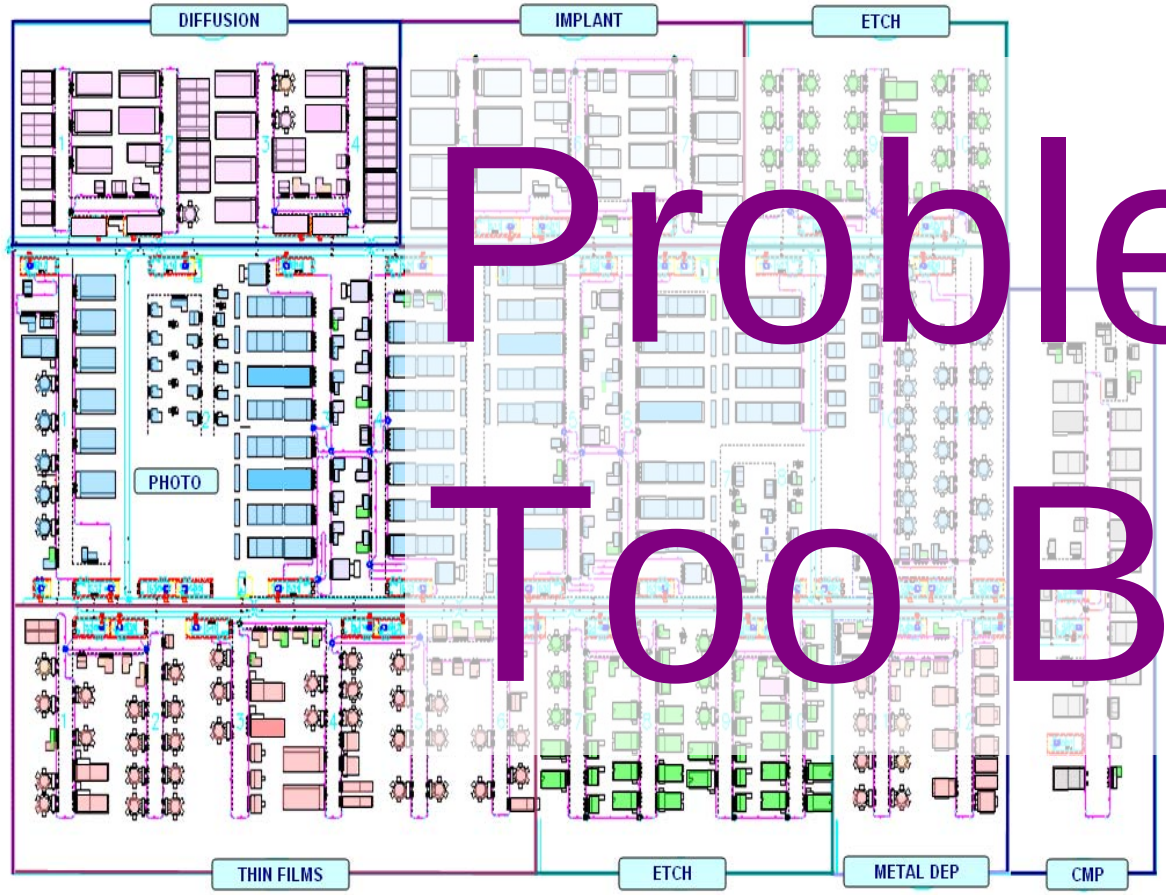
Conclusion: Scheduling is better than Dispatching

Scheduling: Why not Sooner?

- Fab-wide problem is too complex
  - Complex precedence constraints
  - Re-entrant flows
- Optimization was too slow
  - Any computed schedule is out-of-date within minutes
  - Somehow schedules need to be rapidly updated

Solution Approach: Tool Level Scheduling

Generate an optimized schedule in a timely manner



Problem Too Big

- **100s of process tools**
- **100s of process steps**
- **Re-entrant process flows**
- **1000s of lots**
- **Numerous process flows**
- **Optional steps**
- **Variable recipe times**
  - Minutes to hours
- **Variables process sizes**
  - Batch to single wafer
- **Variable transport times**
- **Unpredictable tool failures**
- **Hot lots & Q-Times**
- **Multi-product processing**
- **Reticles**
- **Set-ups**
- **Local policies (RM, phase-in, skip-lots, etc.)**

ILOG Scheduling Solution

ILOG Proposed Solution: Key Ideas

1. One optimization engine for each one of the 6 process area (removes most precedence constraints)
2. Individual optimization engines based upon a detailed tool model and a certain "decomposition":
   - MIP does assignment of lots to tools
   - Constraint programming heuristics produce detailed sequencing and timing
3. Result of optimization: A shift-length (8 – 12 hour horizon) schedule for each tool in the given process area
   - **Schedule is recomputed every 5 minutes!!**
4. Finally: the resulting detailed schedules are used to produce recommended dispatches

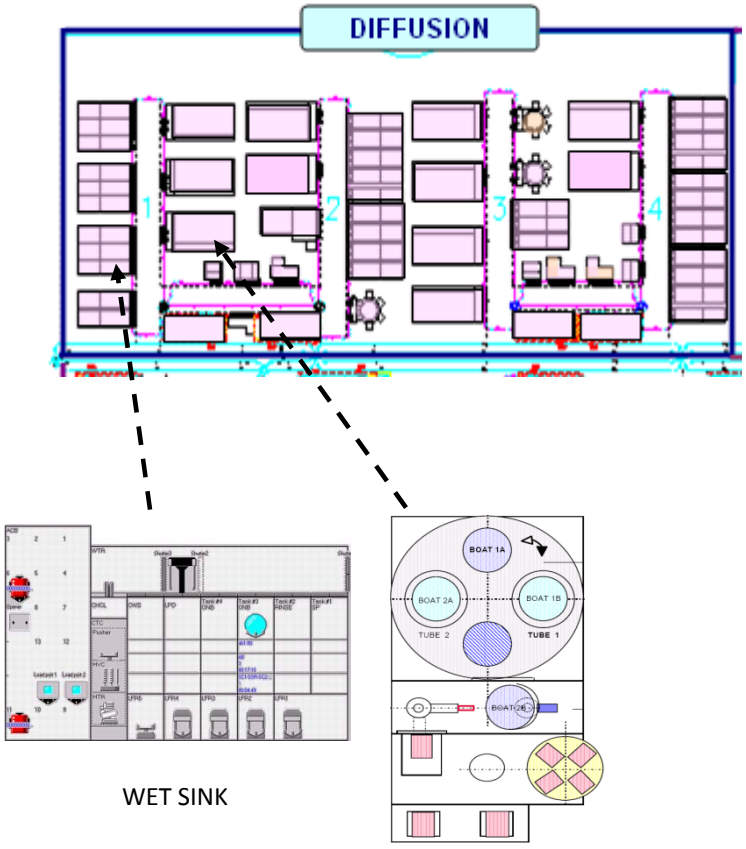
ILOG Solution: One Scheduler for each Area

- Diffusion
  - Rules do a bad job managing batching & process time windows
- Photolithography
  - Most expensive tools: Fab bottleneck
- Etch
- Thin Film
- Chemical Mechanical Polishing (CMP)
- Implant

ILOG Solution: One Scheduler for each Area

- Diffusion (most complex tool set)
  - Rules do a bad job managing batching & process time windows
- Photolithography
  - Most expensive tools: Fab bottleneck
- Etch
- Thin Film
- Chemical Mechanical Polishing (CMP)
- Implant

Diffusion Scheduling Engine



WET SINK

FURNACE

OBJECTIVES

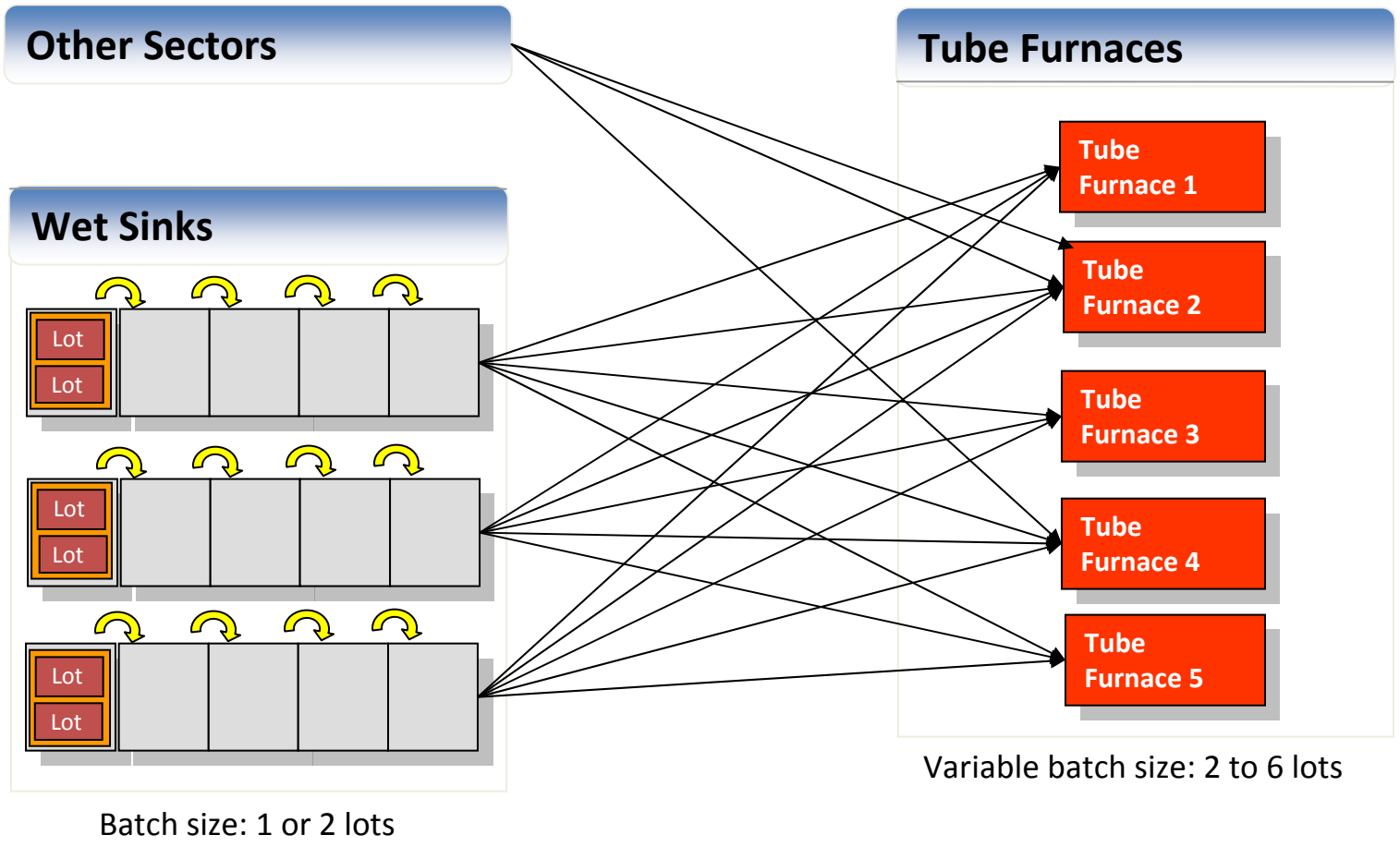
- **Priority weighted throughput**
- **Batch-size weighted throughput**
- **Bay moves**

SOFT CONSTRAINTS

- **Time fence: lots and batches**
- **Urgent lots (Including QTimes)**
- **Training (sequences avoiding setups)**
- **Idle time**

- *HARD CONSTRAINTS*
- **Structural**
- **Tool capacity (time based)**
- **Min <= Batch Size <= Max**
- **Buffer capacity**
- **Wet capacity**

Diffusion Area – Dynamic Batching



Lot Assignment Instance – MIP

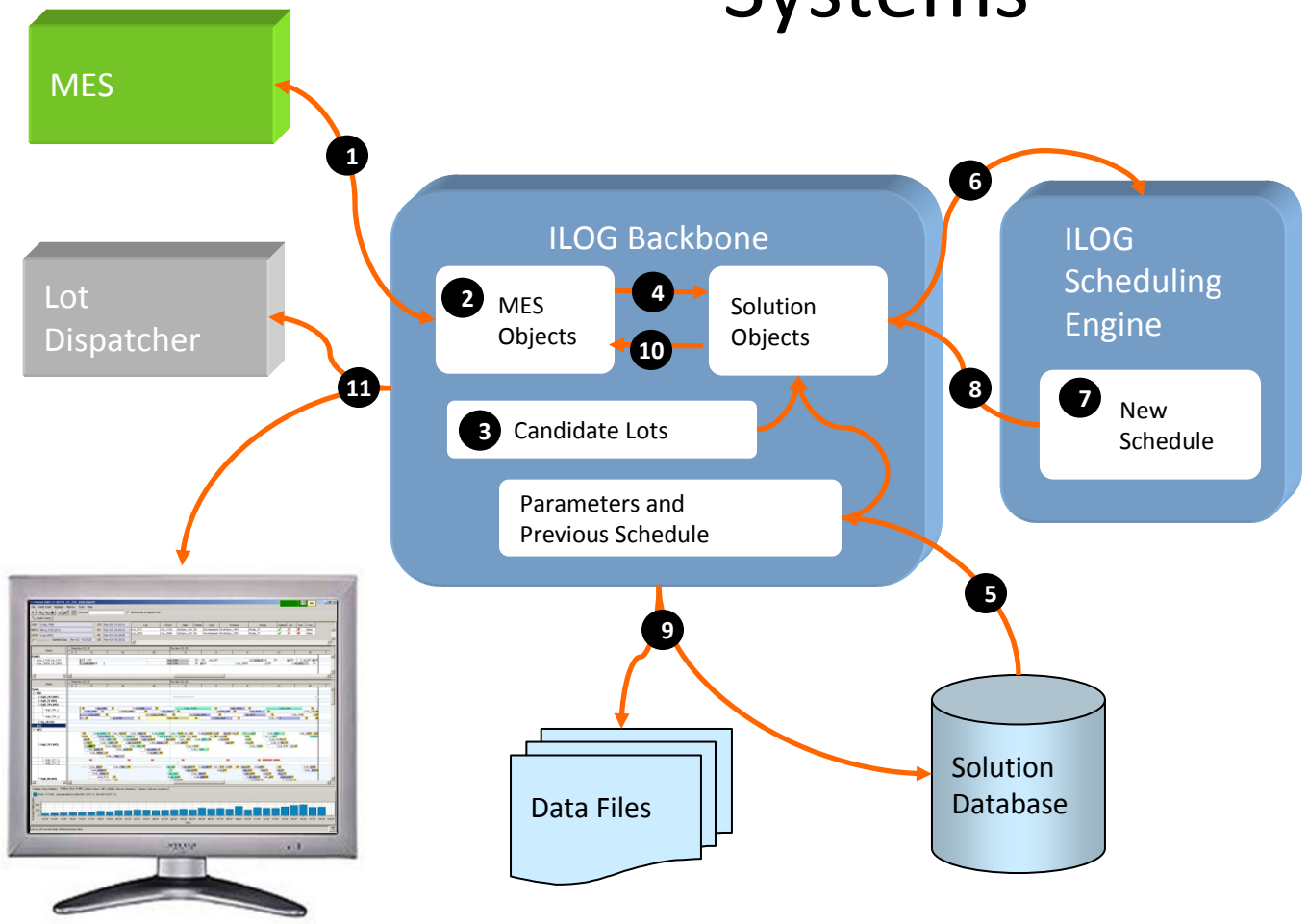
CPLEX 5.0 (1997): 24000 var, 33000 cons, 4000 GIs

```
CPLEX Error  1001: Out of memory.

Error termination, no integer solution.
Current MIP best bound =   -3.9084392492e+02 (gap is infinite)
Solution time = 16520.82 sec.  Iterations = 24359727  Nodes = 854226
```

CPLEX 9.0:

```
   Node  Left Objective  IInf  Best Integer  Best Node    ItCnt   Gap
      0     0  393.2257  1322                393.2257     4853
                366.4625  1185               Cuts: 703     8483  (mostly Gomory cuts)
*   720+  672               0      348.3725   366.3402    28464  5.2%   16 seconds
*  1314+ 1092               0      354.8399   366.3359    43629  3.2%   25 seconds
*  3060+ 2623               0      355.9241   366.2938    94792  2.9%   59 seconds
*  4000+ 2770               0      357.6452   366.2146   127312  2.4%   80 seconds
*  6056  4400               0      357.9718   365.7744   220862  2.2%  137 seconds

Time limit exceeded, integer feasible:  Objective =    3.5797175137e+02
Current MIP best bound =    3.6560278193e+02 (gap = 7.63103, 2.13%)
Solution time =  180.01 sec.  Iterations = 309099  Nodes = 7841 (6124)
```

Data Flow: Integration With Existing IT Systems



1. Load data from MES and other factory systems data sources
2. Data checked
3. Candidate lots selected
4. Data mapped to scheduling engine objects
5. Parameters and previous schedules loaded
6. Engine data checked
7. New schedule created
8. Schedule checked
9. Schedule saved
10. Schedule mapped to MES
11. Schedule converted to dispatch recommendations and published

Running Time:

a.One full cycle takes 5 minutes

b.Schedule computation takes only 20 seconds

Benefits

Improved Fab Performance Metrics

- IBM B323 / Diffusion Area

Results vs. Baseline	FRN	WET
Throughput	8.6%	6.9%
Cycle Time	-25.3%	-8.2%
Hot Lot Cycle Time	-15.4%	-17.9%

Bixby, R., Burda, R., and D. Miller, *Short-Interval Detailed Production Scheduling in 300mm Semiconductor Manufacturing Using Mixed Integer and Constraint Programming*, ASMC 2006.

ROI is substantial

- **Diffusion + Photo achieved Fab-wide 6% cycle time reduction**
  - **Value of 300 mm wafer: $4,000**
  - **Base 20,000 wafers/month throughput and 6% cycle time reduction means 1200/wafers increased throughput**
  - **12 m/y x 1200 w/m x $4,000/w ~= $60M/y revenue**
  - **25-50% profitability/wafer**
  - ROI: $15M-$30M/year

Additional References

- Running in 14 first-tier Fabs in Asia and US
  - 200 mm and 300 mm
    - 300 mm is where the solution brings the most value
  - Types of Fabs
    - Memory
    - TFT/LCD

Other Examples

Other Examples

- ADAC (Konrad-Zuse Zentrum, Berlin)
  - German AAA. 1600 vehicles, 5000 contractors, 20 second response time, installed on 2 of 5 control centers.
- Sabre Trip Shopping (Sabre Decision Technologies)
  - Constraint programming + MIP set covering. 200 millisecond response time for optimization, designed for 6000 optimization threads to co-exist.

Other Examples

- UAV Trajectory Planning (Northrop Grumman)
  - Unmanned Aerial Vehicle obstacle and threat avoidance algorithm. Embedded in real-time operating system. Several hundred variables and constraints, < 1 second solution times.

Conclusion

- This is an exciting time to be an operations-research specialist
  - Data access, model representation, and solution technology advances (the focus of this talk) have enabled whole new application domains
- The emergence of execution-level applications offer the promise of making optimization a mainstream management tool for achieving competitive advantage.

Thank you