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

Robert E. Bixby

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***

Three Success Stories

Using real-time optimization

"Tales from the cutting edge"

Ann Bixby & Brian Downs, Aspen Technologies

"The dance of the thirty-ton trucks"

Martin Durbin, Decisive Analytics
Karla Hoffman, George Mason University

Real-time production-line scheduling for front-end semiconductor fabs

ILOG

Subtext

- **Each of these applications uses optimzation**
  - Linear and Mixed-Integer Programming
- **Question: Did increased solving power really made a difference? Could we have done this 5 years ago?**

Changing the rules of business™

Tales from the Cutting Edge: A real-time capable-to-promise model for meat disaggregation at Swift & Co.

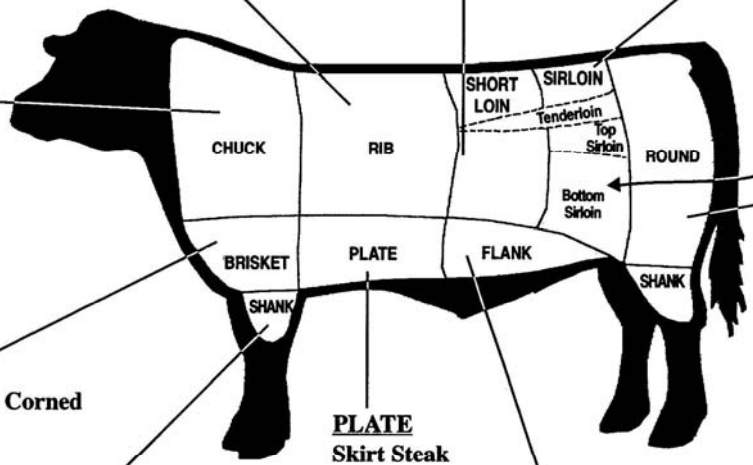
— BEEF CUTS —
Where They Come From

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

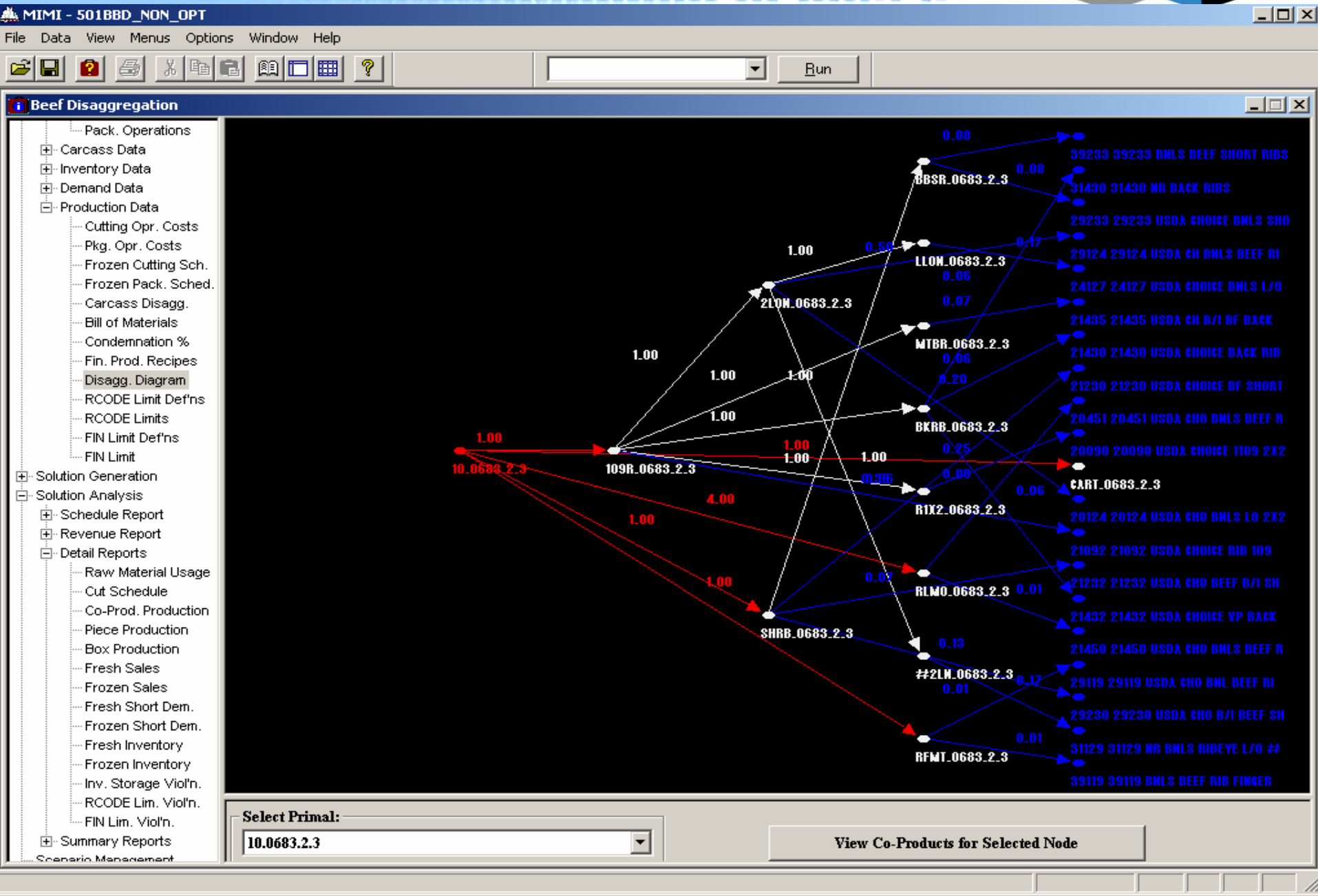
National Cattlemen's Beef Association
444 North Michigan Avenue
Chicago, Illinois 60611
(312) 467-5520

* Beef primals that feature cuts lowest in fat.

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

Changing the rules of business™

The solution

- **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**

An LP Instance from a CTP Model

Size: 241321 constraints; 297517 variables

- **Resolve-time requirement:** **<10 seconds**

- **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

Changing the rules of business™

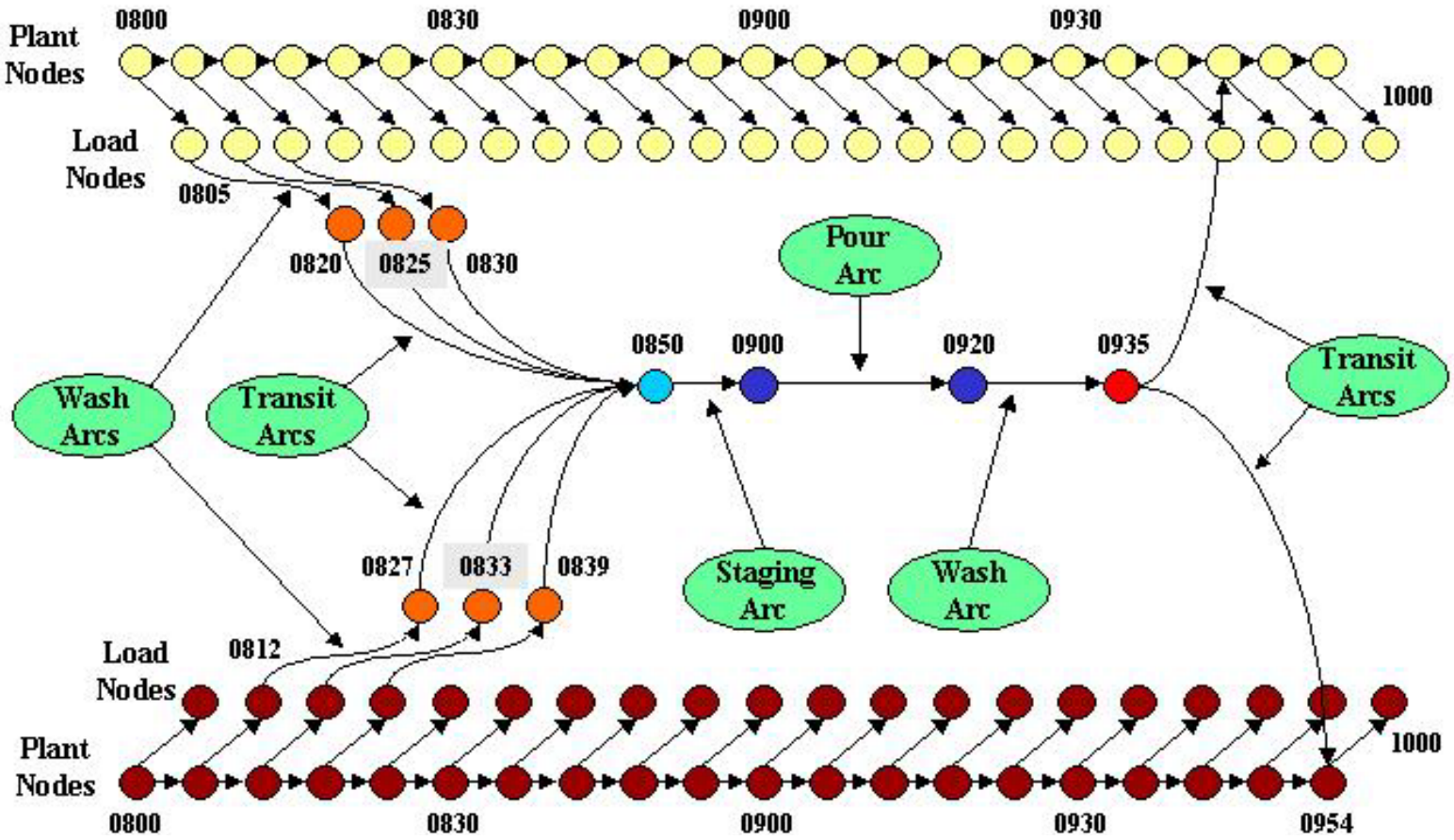
- The Background
  - Virginia Concrete is a part of Florida Rock.
  - They deliver 500-700 loads per day to 150 customers, a total of 5000-6000 cubic yards of concrete per day.
  - Deliveries occur from 10 plants with 125-150 trucks.
- A key characteristic of the business
  - 95% of orders change before 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:

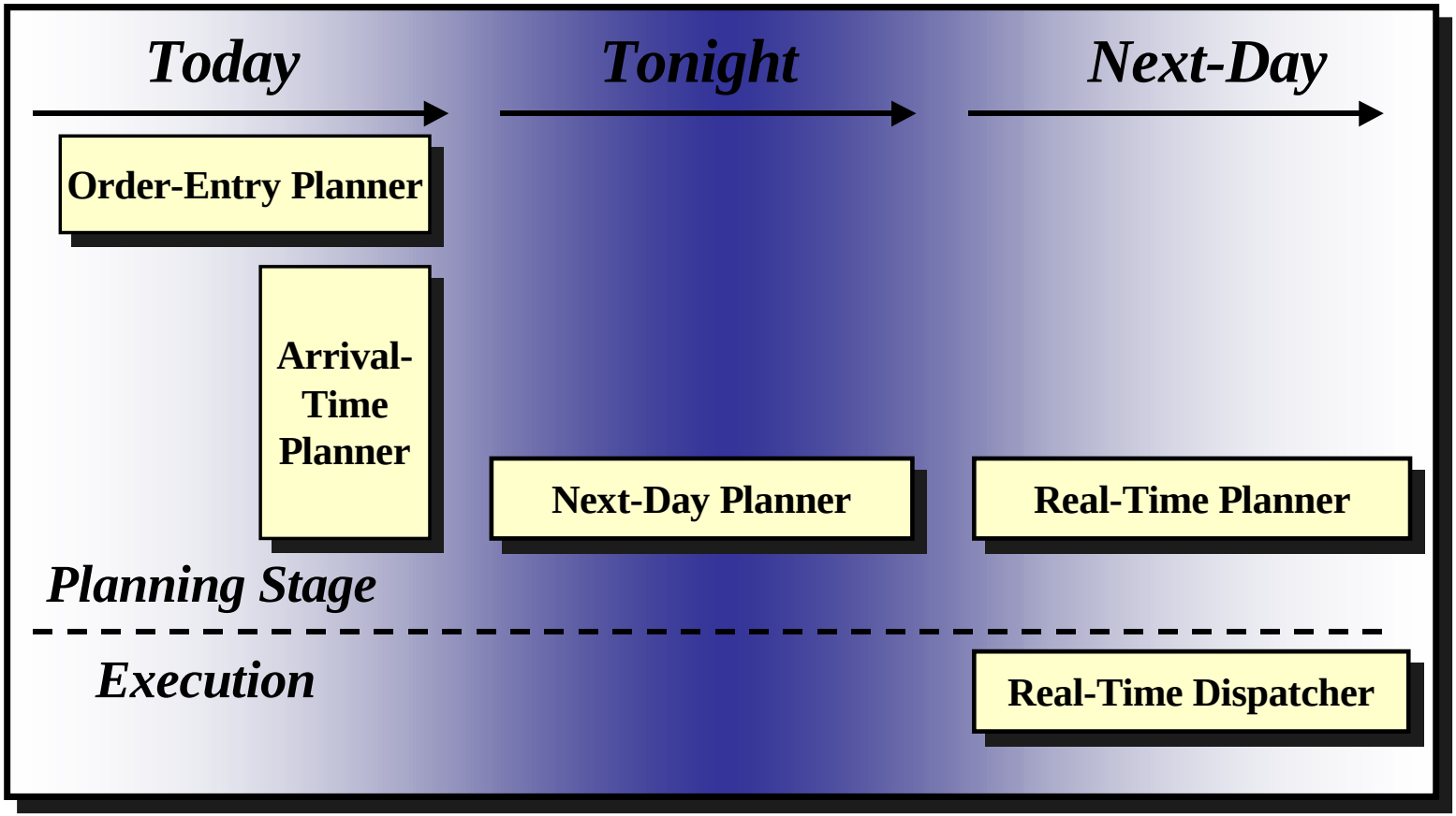
A space-time network



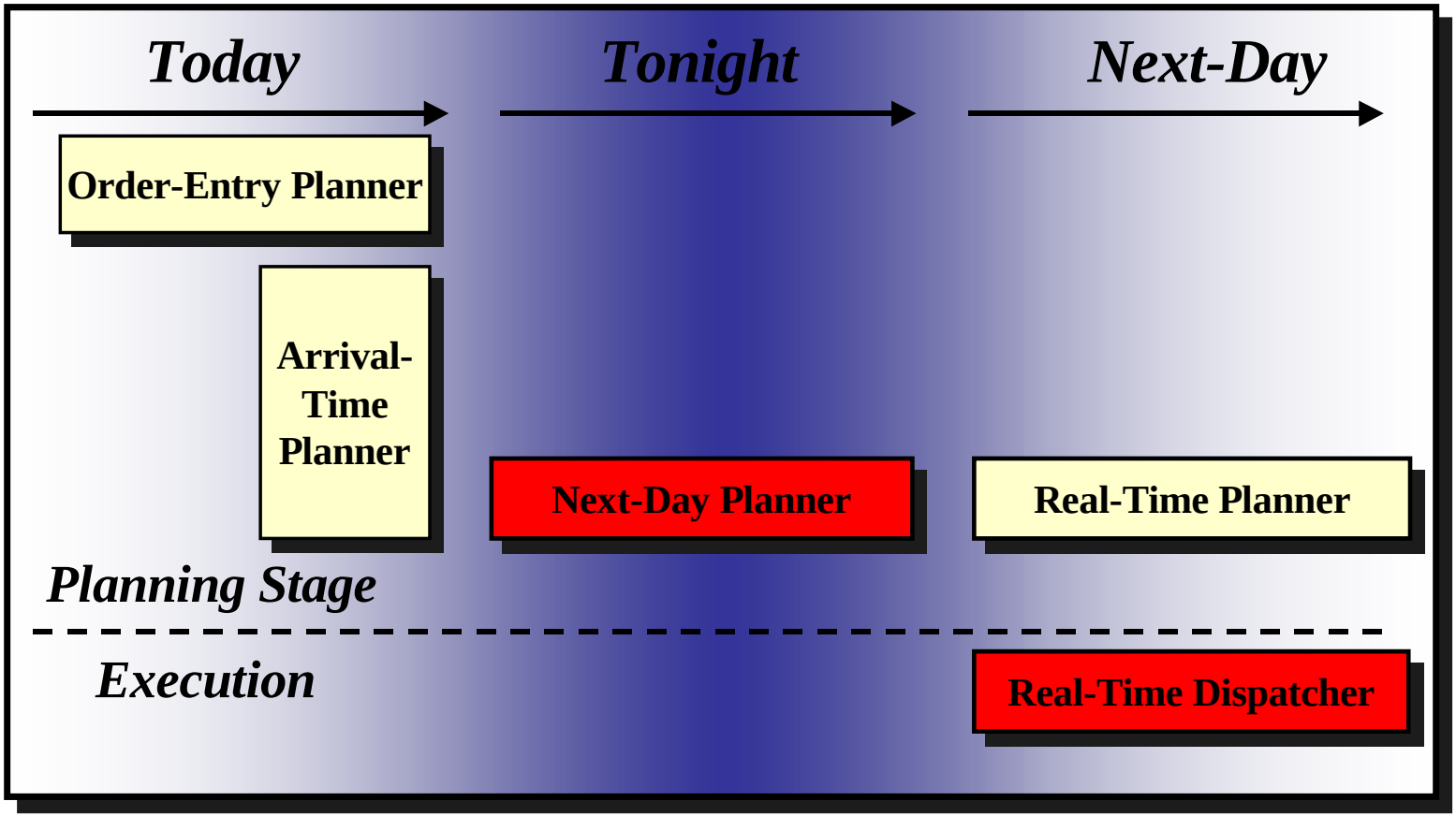
The Optimization Solution

- **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

Changing the rules of business™

- **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

The MIPs

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%

Changing the rules of business™

Fab PowerOps (FPO):

A shop-floor scheduling solution for front-end semiconductor Fabs

Fab PowerOps: Overview

- Semicon background
- The scheduling problem
- Using optimization to schedule fab production
  - The Diffusion module

Semicon Background

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

Semiconductors Today

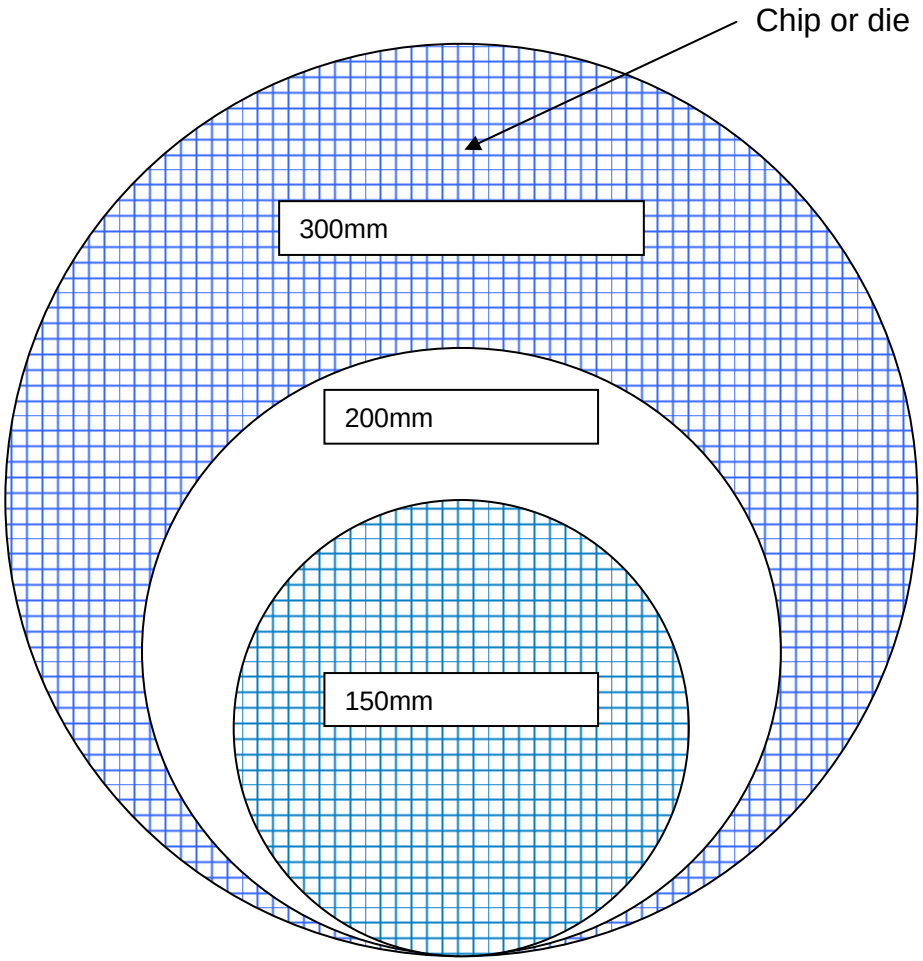
- **Leading the world in manufacturing**
  - **$200+ billion/year business**
  - **$40+ billion/year consumer of other products**
  - **New capital expenditure leader in US manufacturing**
- **Technology driven**
  - **Current generation of Fabs: 300 mm (12 inches)**
  - **Feature sizes down to 65 nanometers**
  - **New Fabs cost $3-4 billion**
- **Tremendous price pressure**
  - **DRAM drops by 10x every 24 months**

The Scheduling Problem

Silicon Wafers

Changing the rules of business™

Wafer size has grown steadily



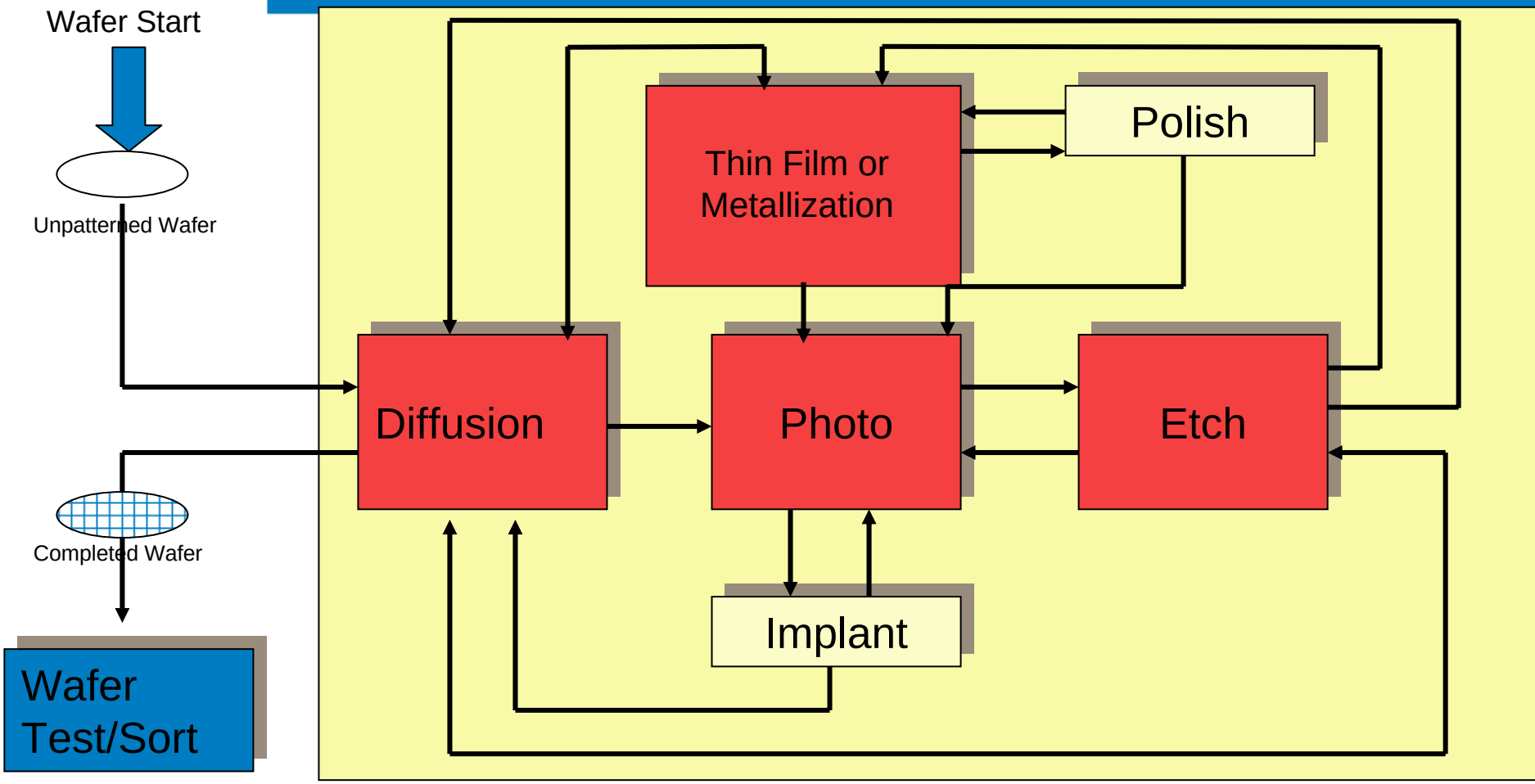
Some facts:

- 300 mm wafers current state-of-the-art
- 500+ chips (dies) per wafer
- Process may require over 300 steps in 30 or more “layers”
- Processing occurs in **lots** of 6-25 wafers (300 mm – held in FOUPs)
- Takes 1-3 months to process a lot

A Cyclic Fabrication Process

Changing the rules of business™

Main front-end processes



The Lot Scheduling Problem

Issues

- Efficient scheduling of lots clearly important
- Scheduling must be flexible and must occur in real-time
  - Processes are delicate
  - Some processing times are unpredictable
  - “Super Hot Lots” disrupt scheduling

Current Scheduling Solutions

Use rules + simulation

- **Real-time dispatching rules are the norm**
  - **Typically involve 30 or more rules**
    - Evolve with time
    - Very complex
    - Order dependent
  - **Must be tuned daily, typically using simulation**
  - **Very brittle and hard to maintain**
    - Can take several weeks to adjust to process changes
- **The solution discussed here is the only optimization-based alternative to rules-based dispatching**
  - **Optimization considered better, but not practical**
  - **Operations people often recognize that “scheduling” is better than “dispatching”**

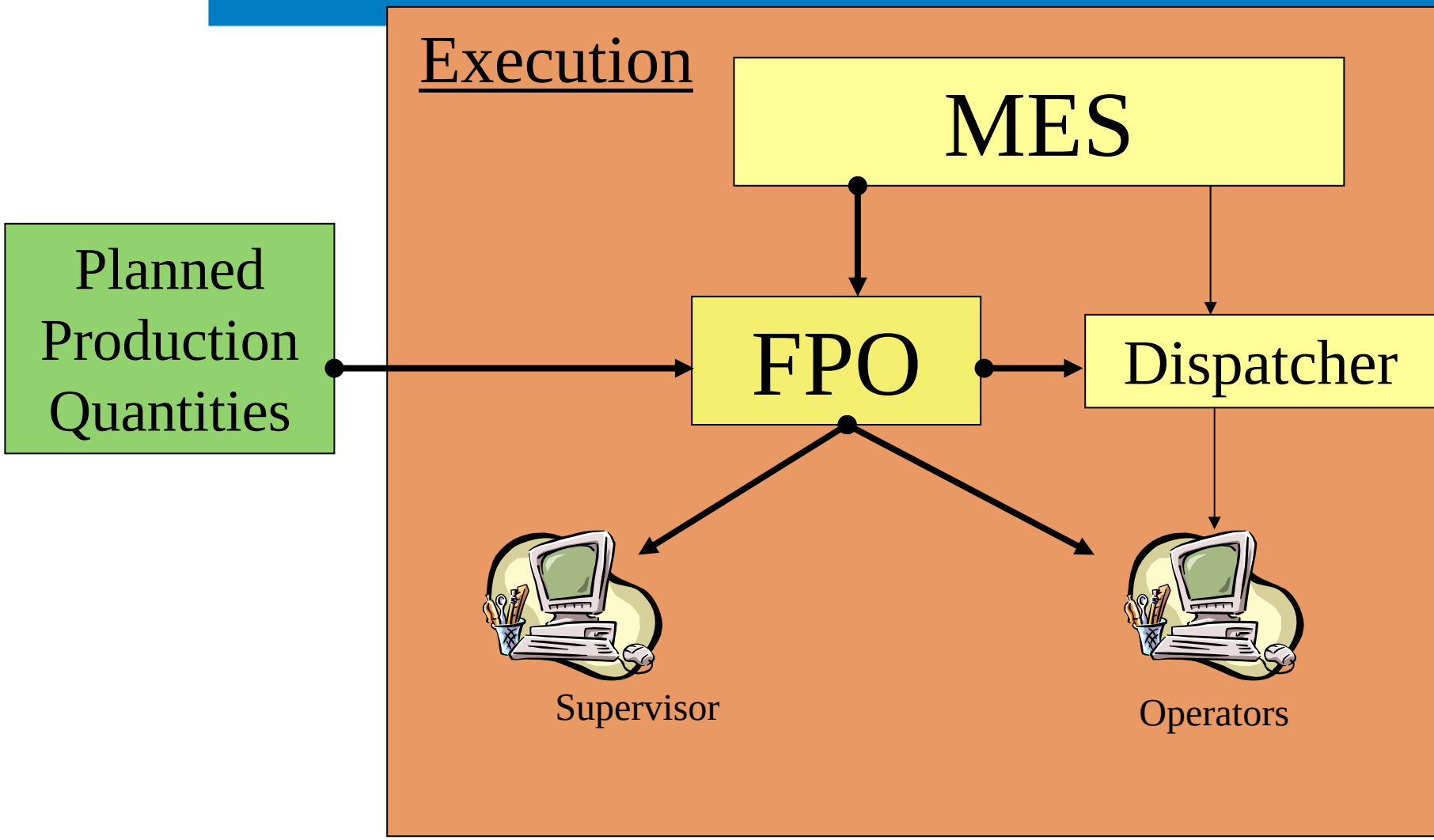
Using Optimization to Schedule FAB Production

FPO modules

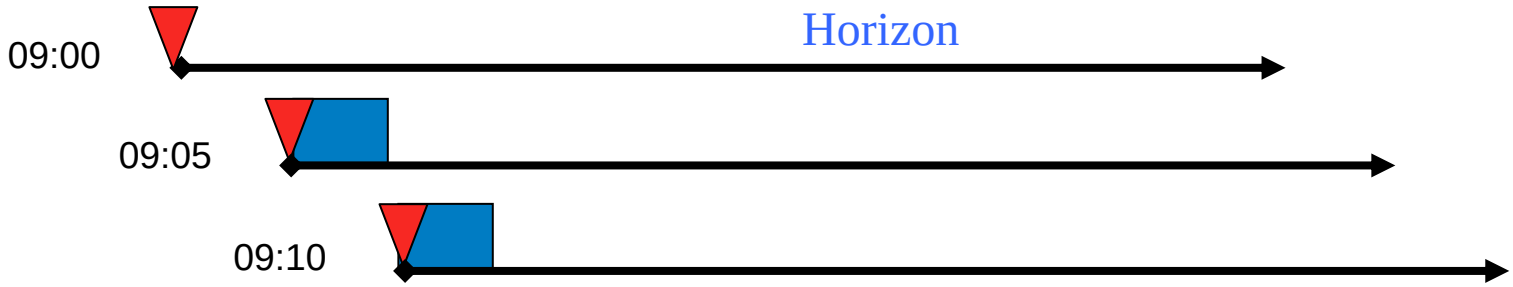
Changing the rules of business™

- **Independent module for each key process**
  - **Photolithography**
  - **Diffusion (clean + tube furnace)**
  - **Etching**
  - **Thin Film (CVD)**
  - **Implant**
  - **Chemical-Mechanical Polishing (CMP)**
- **Optimization model**
  - **Decomposition**
  - **Uses Constraint Programming + Mixed-Integer Programming**
  - **Some constraints are hard, some handled with objectives**
    - Focused on modeling tools rather than processes ➔ translates from one Fab to the next.

FPO fits into the existing software environment



FPO Working Modes (all modules)



- Horizon usually 8 or 12 hours (one shift)
- Rolling horizon
- Runs frequently (every 5 minutes)
- Generates solutions in seconds
  - Mode 1: Beginning of the shift
  - Mode 2: During the shift
- Time fence management

What Enables this Solution?

- Processes have become highly automated
  - Particularly in 300 mm Fabs
- Optimization models are “local” to individual processes, not global to entire Fab
- New decomposition algorithm
  - Took three years to develop
  - Key features exploited: Large # of parallel machines with weak precedence constraints
- Advances in optimization technology

Advances in Optimization

Diffusion Model: Impossible until very recently

CPLEX 5.0 (1998):

```
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 (present):

```
   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)
```

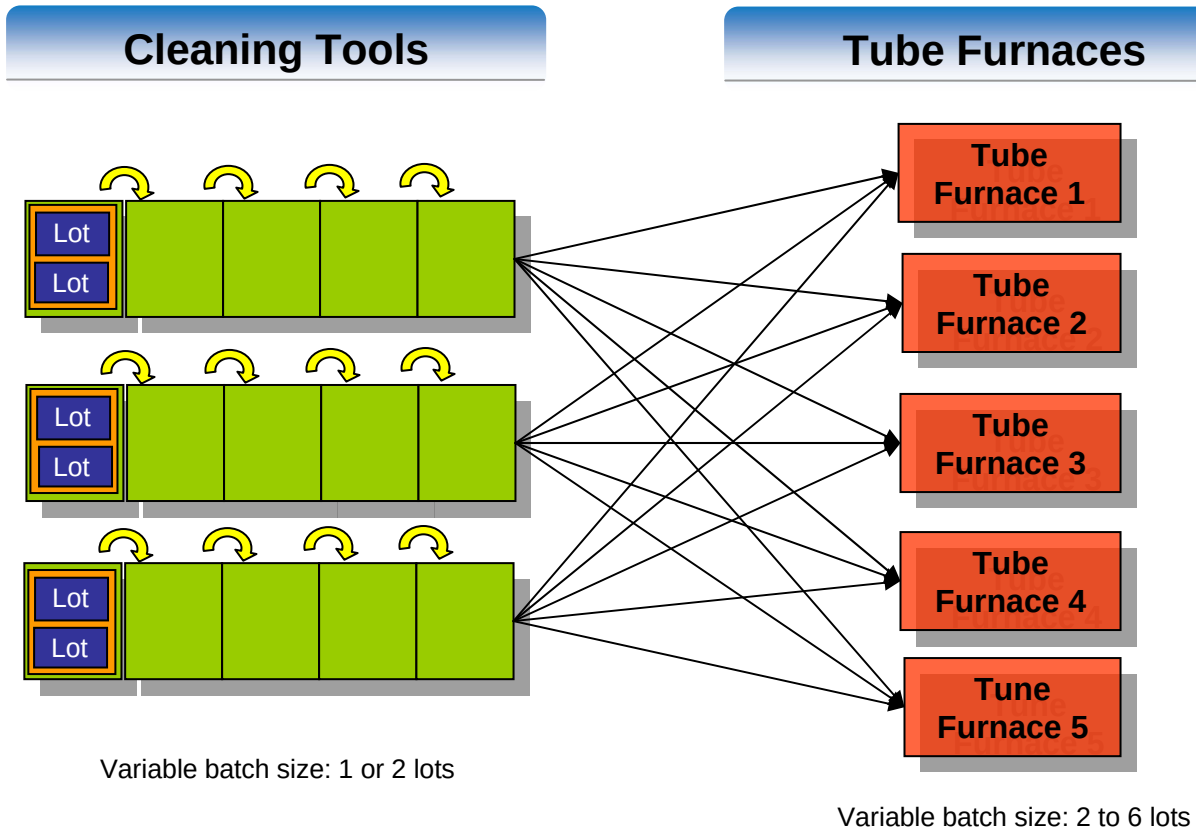
The Diffusion Module

Introduction

ILOG
Changing the rules of business™

Most complex Fab process

- Two processes with dynamic batching and process time windows



Precise Models of Diffusion Process

Process Behavior

- **Tool status**
  - **Production, standby, down**
- **Tool pre-reservation**
- **Tool balancing**
- **Lot processing events and phases**
  - **Operation start, process start, process end, cooling, unloading, operation complete**
- **Lot type, class and priorities**
- **Multiple-lot FOUPs**
  - **Single lot/single recipe, multiple lot/single recipe, multiple lot/multiple recipe**
- **Test steps and scripts**
  - **Measurement, metrology**
- **Inhibits**
  - **Product, route, route/operation, recipe, tool, tool/recipe, process definition/tool, module process definition**
- **Hold mechanism**
- **Phase-in**

Precise Models of Diffusion Process

Tool Behavior

- **Wet tools**
- **Tube furnaces**
- **RTP furnaces**

Precise Models of Diffusion Process

Tool Behavior

- **Wet tool constraints**
  - **Dump-fill constraints**
  - **Picking:** Single wafer, all wafer
  - **Recipe balancing:** Alternate recipes
  - **ML/MR:** Quals
  - **Buffer Management**

Precise Models of Diffusion Process

Tool Behavior

- **Tube furnace constraints**
  - **Batch size constraints:** Tool level, tool recipe level, step/tool recipe level
  - **Process types:** Nitride, poly, copper anneal, cure, standard oxidation, gate oxidation, etc.
  - **Maximum number of FOUPs**
  - **Buffer Management**
  - **BTM (boat transfer module) management**

Precise Models of Diffusion Process

Schedule Objectives

- **(Super) Hot lot management**
- **Throughput**
- **Batch size**
- **Batch time fence**
- **Idle time**
- **Wet capacity**