

Linear Semiconductor Manufacturing Logistics and the Impact on Cycle Time

Peter van der Meulen

BlueShift Technologies, Inc.

3 Riverside Drive, Andover, MA 01810

pvanderm@blueshifttech.com

Abstract:

Fabs need enhanced flexibility to manufacture smaller lots of wafers to reduce cycle time, inventory and WIP, while maintaining equipment throughput, avoiding cross-contamination and ensuring process integrity and yields. Current equipment has increasing difficulty meeting those demands. This paper describes various factors that could lead to optimized choices for the quantity of wafers in a lot of size smaller than 25 wafers, and shows the potential for decreases in cycle time associated with various equipment configurations and wafer lot sizes.

Keywords: Cycle Time, Small Lot, 300mm Prime.

I. Introduction

High mix, low volume manufacturers such as foundries desire smaller lots sizes and single wafer lot tracking. Queuing of small lots (less than 25 wafers), and the rule (left over from 200mm wafer manufacturing) that each wafer has to return to the same slot in the same FOUP, are contradicting each other. Effectively, semiconductor manufacturing is still a batch (of 25 wafers) based operation even on single wafer tools. Several other tools in the fab have batch behavior, such as Vertical Furnaces and Wetbenches. Ion implanters sometimes use batch load locks, which also creates an affinity for a 25 wafer lot size. Table 1 shows a modern fab tool list and the behavior associate with typical tools.

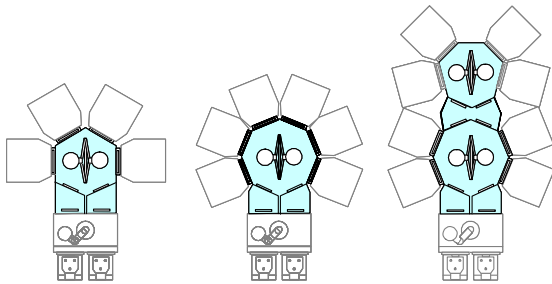


Figure 1. Traditional Clustertools.

In traditional cluster tools (Figure 1), a monolithic block serves as a central wafer handling system. Such an approach suffers from serious drawbacks. First, wafers have to be handled many times by the same robot and have to be returned to the same slot in the same FOUP. The last wafer in the lot has to be returned to the FOUP before a new lot can be started for cross-contamination purposes. This rule forces the AMHS system to perform a fast FOUP exchange on the tools' load port. An effect that is certainly made worse if less than 25 wafers are in the FOUP. Second, there is no obvious place for an integrated inspection module. It is very difficult to perform an inspection in between process modules: one either has to halt the robot or to add an integrated metrology module on a process facet. Third, vacuum isolation (to prevent cross contamination between for example PVD and CVD chambers) has resulted in unwieldy central vacuum handlers that necessitate a physical separation between the front (CVD) chambers and the back (PVD) chambers. Such an arrangement can result in poorly matched throughput between chambers and thus a poor Cost of Ownership performance of the tool (Figure 2).

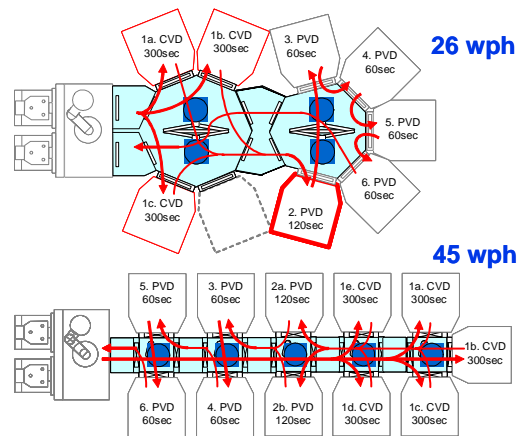


Figure 2. Bottleneck Process Chambers

The net effect of the above drawbacks is a reduced ability for fast cycle times of wafers lots, which in turn leads to higher WIP levels and to higher wafer inventories.

Typical Tool List	Subcategory	Supplier	Tool Type	Model Name	Affinity	Behavior	#Tools
Ion Implanters	High Current Implant	VSEA	Single Wafer, Batch Load Lock	VIISta HCP	25	Steady State	10
	Medium Current and High Energy Ir	VSEA/ACLS	Single Wafer, Batch Load Lock	VIISta 810	25	Steady State	11
RTP	Implant Anneal	AMAT	Direct Load EFEM	Vantage	2	Steady State/Periodic	15
	Oxidation	ASMI	Vertical Furnace	A412 Smartbatch	25	Batches	3
EPI	Epitaxial Silicon	AMAT	Clustertool	Centura EPI	1	Steady State/Periodic	6
Metals	Liner/Barrier	AMAT	Double Clustertool	Endura iLB PVD/CVD	1	Steady State/Periodic	11
	Tungsten	NVLS	Lazy Suzan Clustertool	Altus	1	Steady State/Periodic	10
	PVD Copper Barrier/Seed	AMAT	Double Clustertool	Endura CuBS PVD	1	Steady State/Periodic	10
	Silicides	AMAT	Double Clustertool	Endura ALPS Ni PVD	1	Steady State/Periodic	10
Dielectrics	LPCVD	Kokusai	Vertical Furnace	Quixace	25	Batches	4
	PECVD	NVLS	Lazy Suzan Clustertool	Vector	1	Steady State/Periodic	20
	Low k (PECVD)	ASMI	Vertical Furnace	A412 Smartbatch	25	Batches	4
	High Density Plasma	AMAT	Dual Wafer Clustertool	Producer APF	2	Mini Batches	23
ECD	ECD Plating	NVLS	Multistation Plating Tool	Sabre	2	Mini Batches	6
Etch	Silicon Etch	Hitachi/LRCX	Cluster/4PM	U-8250	1	Steady State/Periodic	56
	Dielectric Etch	TEL/LRCX	Cluster/4PM	Tellius	1	Steady State/Periodic	17
Dry Strip	Photoresist Removal	MTSN/ACLS	Dual Wafer Clustertool	Suprema	2	Mini Batches	20
CMP	Copper CMP	AMAT	4 Head Polisher	Reflection	4	Mini Batches	11
	Dielectric/Poly/Tungsten CMP	AMAT	4 Head Polisher	Reflection	4	Mini Batches	11
Lithography	Lithography	ASML	Stepper	Twinscan	2	Steady State	46
Wet Clean	FEOL/BEOL Clean	DNS	Wetbench	FC3100	25	Batches	9
Track	FEOL/BEOL	TEL	Track	Lithius	1	Steady State	46
TOTAL TOOLS (Not including metrology and inspection)							359

Table 1. Modern Equipment List

However, the continuing trend for smaller lots and thinner layers as well as the trend for more integrated processing (such as with additional liner or barrier layers or etch stops integrated with a main deposition) requires a much more flexible tool architecture that allows for small lot sizes to be easily processed without putting undue burden onto the AMHS system, and without many lot split and merge operations. We are proposing a novel architecture for next generation process equipment that avoids the limitations of existing equipment architectures, while being significantly more cost effective and flexible.

II. Linear Manufacturing

Fab cycle time is generally given by the following simplified formula [1]:

$$\text{Cycle Time} = \text{Sum of Process Times} + \text{Sum of AMHS Move Times} + \text{Sum of Stocker Wait Times} \quad (1)$$

Generally the Wait Time component is about 2-4x larger than the other 2 components [2]. This is partly because tools are not always immediately available and partly because tools sometimes accumulate lots before processing (as is the case with Wetbenches and Vertical Furnaces). Reducing the lot size has a positive effect on both: effects: tools are not occupied for very long when processing a small lot and become available sooner, and if smaller lots are processed, in stead of large lots, there is less time needed to accumulate all the wafers for a load. However, reducing the size of the lots that are used in the fab (assuming that the carrier holds exactly one lot), results in an increase in AMHS traffic, which at the extreme can cause it to become a bottle neck itself.

An alternate approach to the current tool set is to implement a linkable, linear vacuum handling system

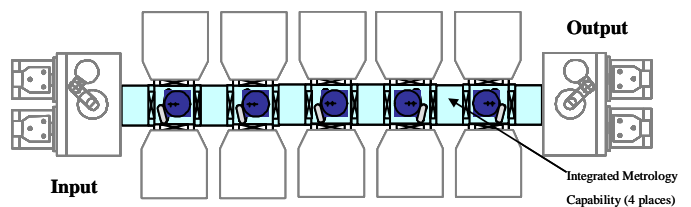


Figure 3. Linkable, linear process tool

(Figure 3). Several advantages immediately emerge: a significant reduction in footprint, a lower capital cost, the ability to provide many levels of vacuum isolation allowing integration of many different process types, and the ability to stack lots back to back, which results in a significantly higher system throughput. The central linkable element can be expanded, even after the original equipment has been installed.

Other industries have long since moved to linear manufacturing methods. Linear operations are significantly more effective for high throughput, well controlled operations. Figure 4 shows the difference in throughput, which was modeled for a serial and parallel flow in 6 process chambers. It is clear from the figure that for short process times the linear architecture has superior throughput. From formula 1 it can be seen that increasing throughput is equivalent to reducing cycle time. In a parallel flow, the first robot in the first link has to handle every wafer if wafers are returned to the same location. But in a linear system, the wafers do not have to come back, which results in a better throughput for robot limited flows. Conversely in a serial flow, the central robot in a cluster tool has to handle every wafer and becomes a limiter for short process times (Figure 4). In a linear tool the throughput stays significantly higher partly because wafers do not have to come back and partly because there are multiple robots doing the handling.

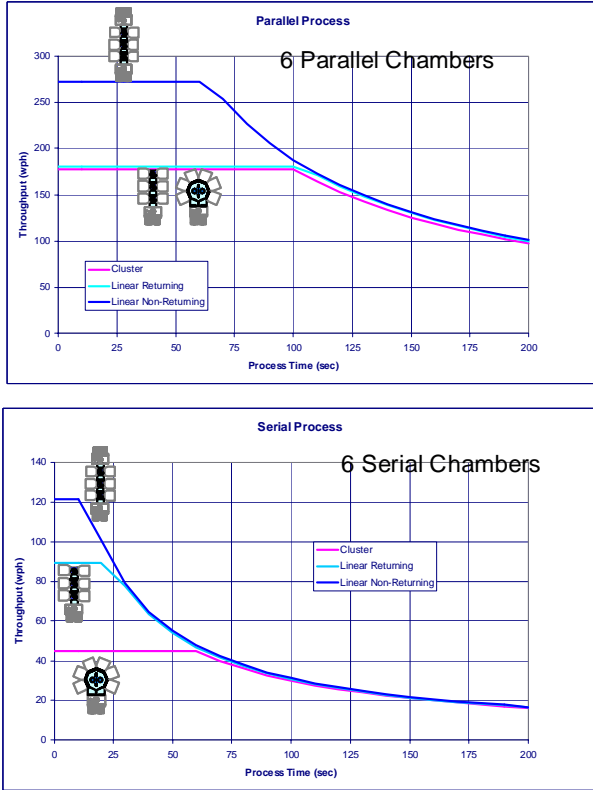


Figure 4. Throughput comparison in linear and cluster tool configurations

Note 1. A serial flow in a linear tool can have constant throughput even if additional process steps are added, which is one of the advantages of linear architectures: it is possible to add additional steps without impacting throughput.

Note 2. If a flow is process limited, there is little difference in throughput, since the process time is dominant. However, in a linkable system one can always add links to increase throughput to the point at which the flow becomes robot limited.

Note 3. Process times get shorter when films get thinner for smaller device geometries. This makes linear architectures more favorable, since throughputs are generally higher for short process times.

If done correctly, linear manufacturing can also be much more effective for small lot sizes, since “queue fill” and “queue empty” effects can be minimized. In figure 5, we show the effect of a tool being filled in the dynamic throughput, which is defined as the elapsed time divided by the number of wafers returned to the carrier. Ultimately the dynamic throughput will approach the steady state

throughput, but because in this example we used 4 process chambers and a 25 wafer carrier, there will be a time when only one process chamber is in use, which is a normal operating method in fabs where one tries to avoid cross-contamination between lots.

III. Cycle Time Reduction and Small Lots

The model in figure 5 shows fairly typical: cluster tool behavior. The system runs in a process limited mode, not in robot limited mode, and typically has a number of chambers that is not an even divider into the number of wafers in the carrier. As a result, cycle time in the fab could be reduced by adding more process chambers to process-limited tools (which increases throughput), or by matching the lot size to the number of process chambers so they divide evenly. Unfortunately, cluster tools are limited by the number of process facets, so adding process modules is mostly not an option. A linkable, linear approach circumvents those weaknesses: it is always possible to add additional process chambers, and it is relatively easy (for small lots) to match the number of wafers in the carrier to the number of process stations in the tool. Furthermore, a linear tool can have a significantly higher throughput, since wafers do not have to be moved back down the line, which frees up the handling robots.

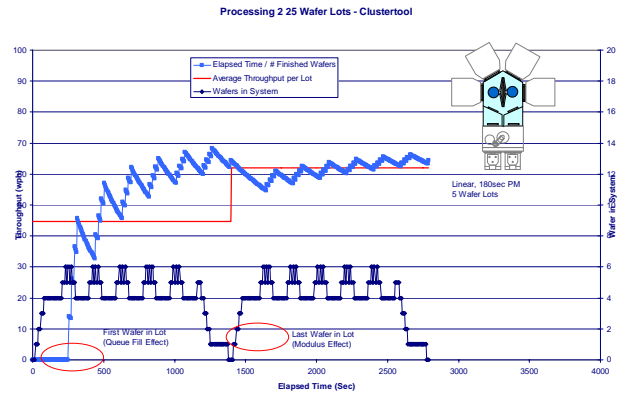


Figure 5. Queue Fill and Queue Empty Effects

Figure 6 shows the total processing time for 2 lots of 25 wafers. The clustertool has a slightly longer finishing time than the linear tool (orange curve) because of the tool being only partially full at the end of the first lot. This effect is exacerbated when going to for example 5 wafer lots (red curve). A linear tool however processes faster (blue curve) even for small lots since it is able to maintain lot separation inside the tool.

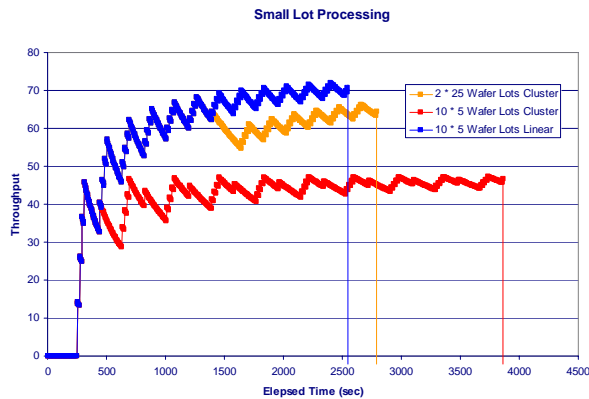


Figure 6. 50 Wafer finishing time, Cluster vs. Linear

Scheduling software however, tries to schedule wafers for optimum throughput. A tool that processes wafers in parallel modules will therefore sometimes exhibit behavior as modeled in figure 7 that is periodic in nature. In other words, the time interval between subsequent wafers returned to the carrier varies. Three wafers come back 20s apart and the fourth wafer comes back 130s after wafer three. Two effects occur when process locations are added (as can be done on a linear system): the throughput increases (which reduces cycle time) and the arrival time variability is reduced.

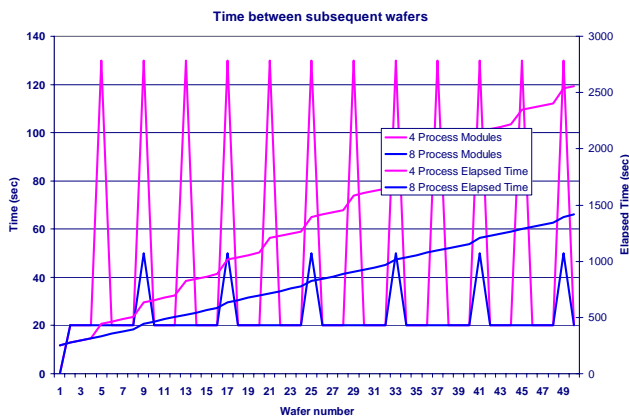


Figure 7. Periodic Behavior in Tools

IV. Mini Batch Tools

The above periodicity effect led us to look at the periodic behavior or affinity of a typical fab toolset (Table 1). Today's tools are mostly single wafers tools, and although they exhibit the periodic behavior in figure 7, they have a primary single wafer affinity. However, recently a new class of process tools has come to market that process wafers in mini-batches of 2, 3, 4 or 5 wafers to more effectively compete with vertical furnaces and single chamber cluster tools. Examples of these tools are the Applied Materials Producer™ (Figure 8), the Mattson Suprema™ and the Jusung Cyclone™ ALD.

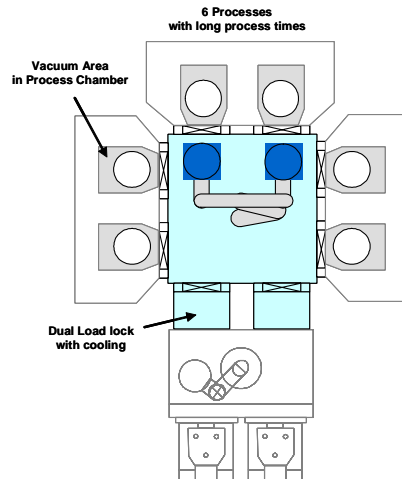


Figure 8. Sketch of Applied's Producer™

The Applied and Mattson tools process 2 wafers simultaneously, the Jusung tool processes 5. Other companies are working on 3 and 4 wafer process chambers such as in figure 9. The Novellus Vector™ uses a "Lazy Suzan" style process chamber and thus behaves more like a single wafer tool than a mini-batch tool. CMP tools sometimes use 3 or 4 polishing heads in parallel or in series depending on the recipe and architecture, and thus have a 1, 3 or 4 wafer affinity. This leads us to conclude that whatever the choice of smaller lot size will be for the transition to 300mm Prime, it should take into account the affinity for tools to certain wafer multiples.

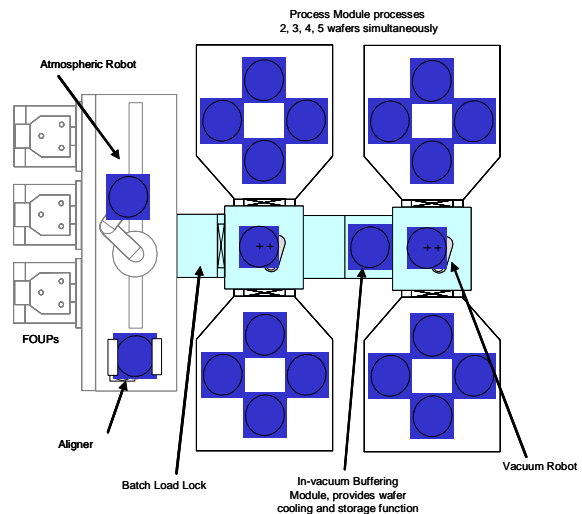


Figure 9. Mini-batch Process Tool Example

Number of Wafers in Carrier		
AMHS Factor	Pro's	Con's
1	25.0 Simple	Too many AMHS moves, Not good for "Producer" style tools
2	12.5 Simple, even, matches "Producer" style tools	Many AMHS moves
3	8.3 Small 3 wfr mini-batch process chambers	Odd number, matches very few tools
4	6.3 Even number, 4 wfr mini-batches, "Producer", SEMI Slot Valve	6x AMHS moves
5	5.0 Compatible with 25 wfr FOUPs	Odd number, matches very few tools, Uneven tool behavior
6	4.2 Even number, "Producer"	No current mini-batch tools with 6 wafers
7	3.6	Terrible choice: prime
8	3.1 Even number, 4 wfr mini-batches, "Producer"	3x AMHS moves
9	2.8	Only divisible by 3
10	2.5 More compatible with 25 wfr FOUPs, divisible by 2	Maybe not enough benefit
11	2.3	Terrible choice: prime
12	2.1 Even number, 4 wfr mini-batches, "Producer"	Not enough benefit
24	1.0 Matches better with many tools, no changes	

Table 2. How small should a small lot be?

Whichever the choice of number of wafers is, it would make sense to have a fab's mini-lot size set in proportion to the tool's mini-batch size. In figure 10, we display the affinity to a particular number of wafers in a fab and estimate how this might change over the next few years.

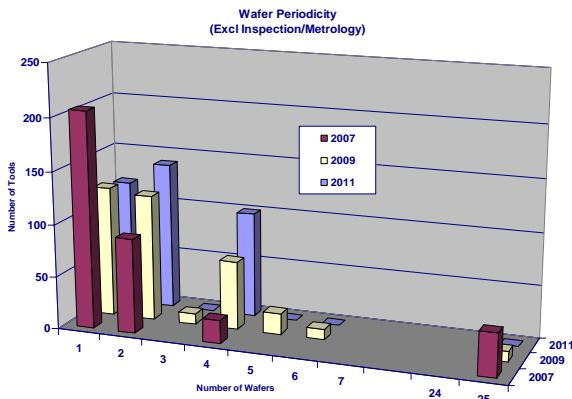


Figure 10. Wafer affinity and projection to the future.

Mini-batch tools need to load the wafers into a vacuum system through a load lock. Wafer heating and cooling has to be done in a mini-batch fashion as well, which can sometimes be problematic since not enough locations may be available on the tool to simultaneously heat or cool a multitude of wafers. The SEMI standard for (process chamber) slot valves (SEMI E21.1-1296) gives a vertical opening height of 50mm, which lends itself well to load 4 wafers into a load lock (figure 11). Although only a minor point, it seems that a 4 wafers lot size easily matches with the SEMI standard slot valve height. Certainly 5 or 6 wafers are much harder to accommodate inside a standard valve height. The advantage of linear tools is also apparent for mini-batch process chambers: typically heating and cooling are done at the beginning and end of a process flow, which is easier in a linear system because in a cluster tool one has to combine the heating and cooling

in a single load lock module, since the return path overlaps with the entry path, whereas in a linear tool architecture the heating and cooling functions can be optimized at the entry and exit of the tool.

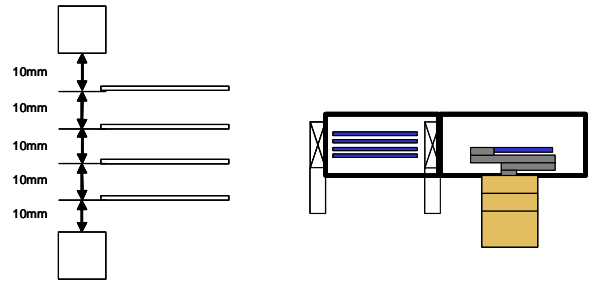


Figure 11. Wafer Planes in Load Lock using E21 valve.

V. Process and Yield Control

An indirect contributor to fab cycle time is rework and yield loss. Obviously rework leads to longer cycle times. Yield losses (which sometimes constitute an entire lot) can be significant in traditional architectures. Integrated metrology is typically added to the EFEM (Equipment Front End Module) of a cluster tool. However, by the time a first wafer is returned to the EFEM, a significant portion of the lot is already in the tool. If a serious problem is detected at that stage, one may have to scrap the most of the lot. In linear architectures integrated sensors can be deployed in between the links (Figure 13), which allows for measurements to be made without impacting throughput. If a problem is detected only a few wafers in the lot are affected, and recovery is easier: wafers already past the problem point can continue to process, and wafers that have not reached the problem point can be returned to the source for later processing, possibly on a different tool. The same location in-between links could be employed for in-process wafer cleaning, which could significantly contribute to yield improvements.

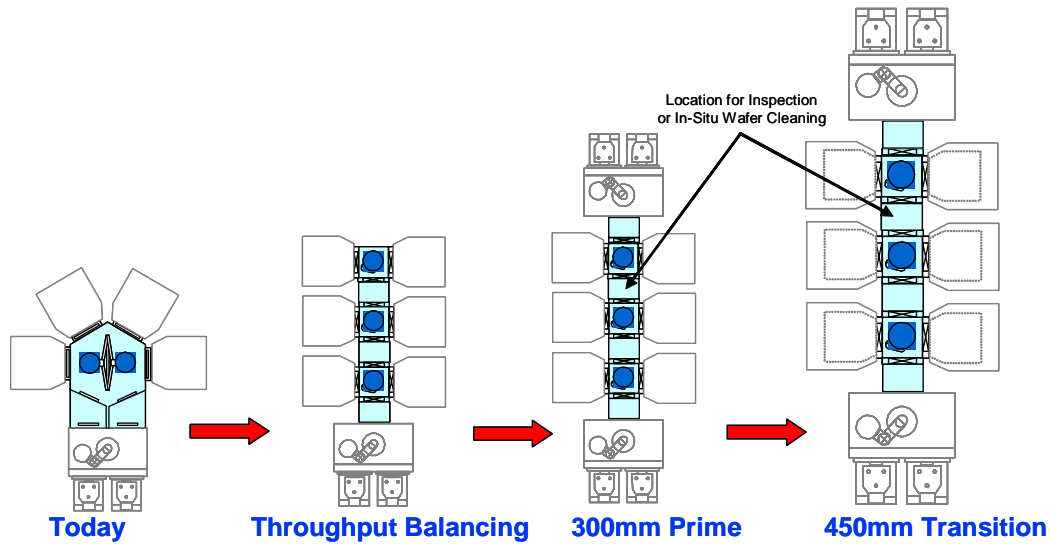


Figure 12. Progression for 300mm Prime and 450mm also showing inspection or cleaning locations.

In general, keeping wafers in vacuum longer could be beneficial to wafer yield. Every pump/vent cycle can potentially add particles to the wafer. Furthermore, every time the wafer leaves a tool it has to be cooled and then is frequently re-heated in the next tool, which wastes energy, adds non-value added time to the process flow (increasing cycle time) and adds to the thermal budget of the wafer.

VI. Conclusions and Recommendations

Cycle time in fabs can be reduced by increasing equipment throughput, which can be readily accomplished in linear, linkable system architectures. Throughput balancing (running equipment at the automation limit rather than the process limit) is an easy first step (Figure 12). Small lot manufacturing on linear systems leads to higher throughputs for short process times, all the while maintaining full separation between lots and avoiding cross-contamination. Eventually linear systems can ease the transition to 450mm wafers since the footprint scaling is more linear than the scaling of a cluster tool (Figure 12). The above discussion leads us to conclude that the following should be considered for the implementation of short cycle time, small lot manufacturing fabs:

1. Drop the requirement for wafers to return to the same slot in the same carrier.
2. Pick an even number of wafers in the carrier.
3. Take into account that many systems have an affinity for 1, 2, 4 or 6 wafers and that the SEMI slot valve standard fits 4 wafers nicely.
4. Linear systems could significantly boost throughput in two important ways: throughputs are higher because wafers are not returning, and process limits can be avoided.
5. Wafers could be kept under vacuum longer, reducing cross-contamination, reducing thermal effects and reducing cycle time.

References:

- [1] Oliver Rose, Proceedings of the SMOMS'99 conference (1999 WMC), 1999.
- [2] Samuel C. Wood, "Factory Modeling", Handbook of Semiconductor Manufacturing Technology, McGraw Hill, 2000, p1103.
- [3] Ken van Antwerp, "Automation in a Semiconductor Fab", Semiconductor International, Dec 2004.