EDWARDS

Cooling and Particulate Challenges

For Next Generation, Atomic Layer Etch Technologies

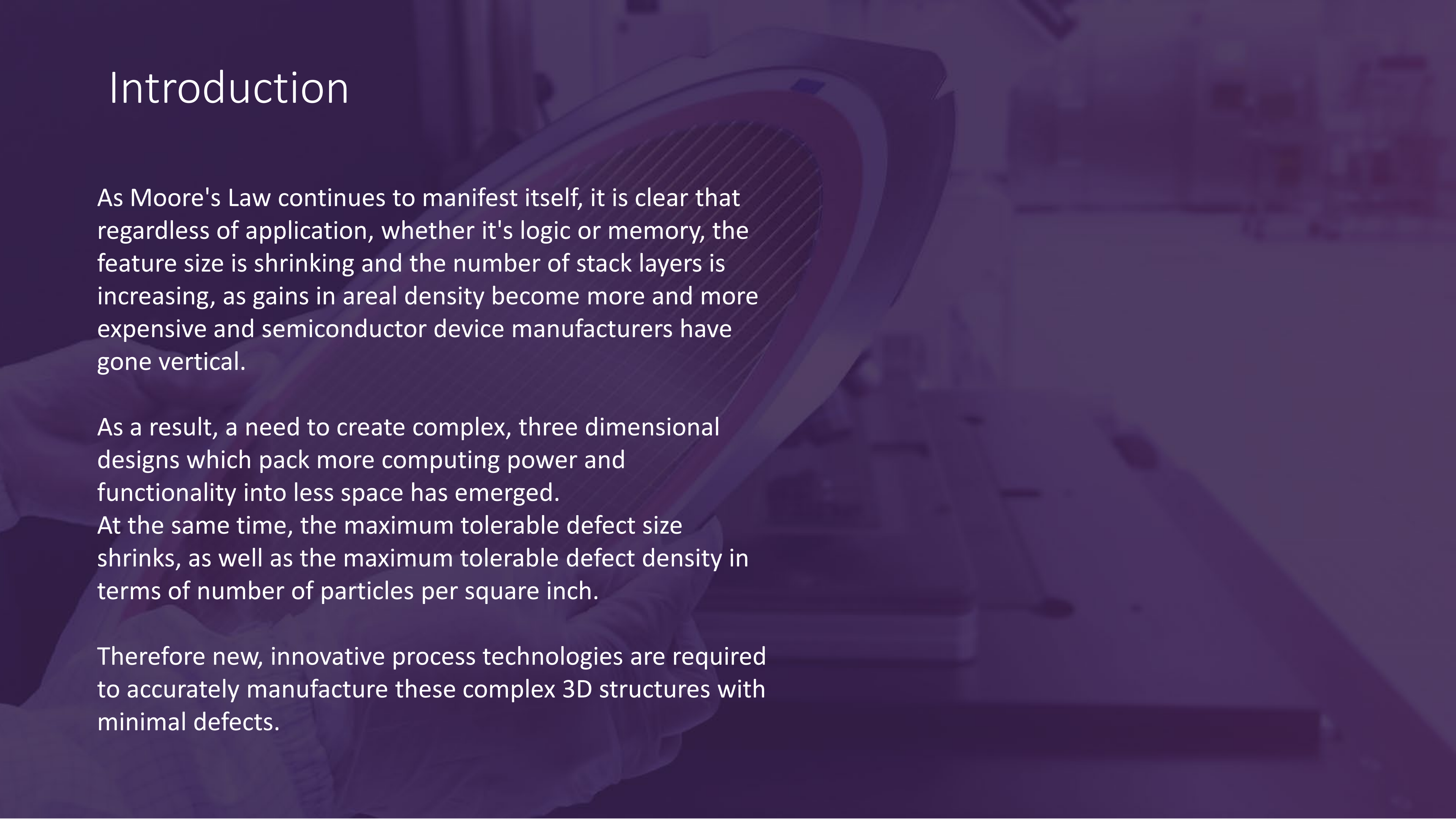
Raj Melkote
VP of Engineering

Declan Scanlan
Market Sector Manager

Dawn Stephenson
Business Development Manager

Kaustubh Kale
Refrigeration Technology Engineer

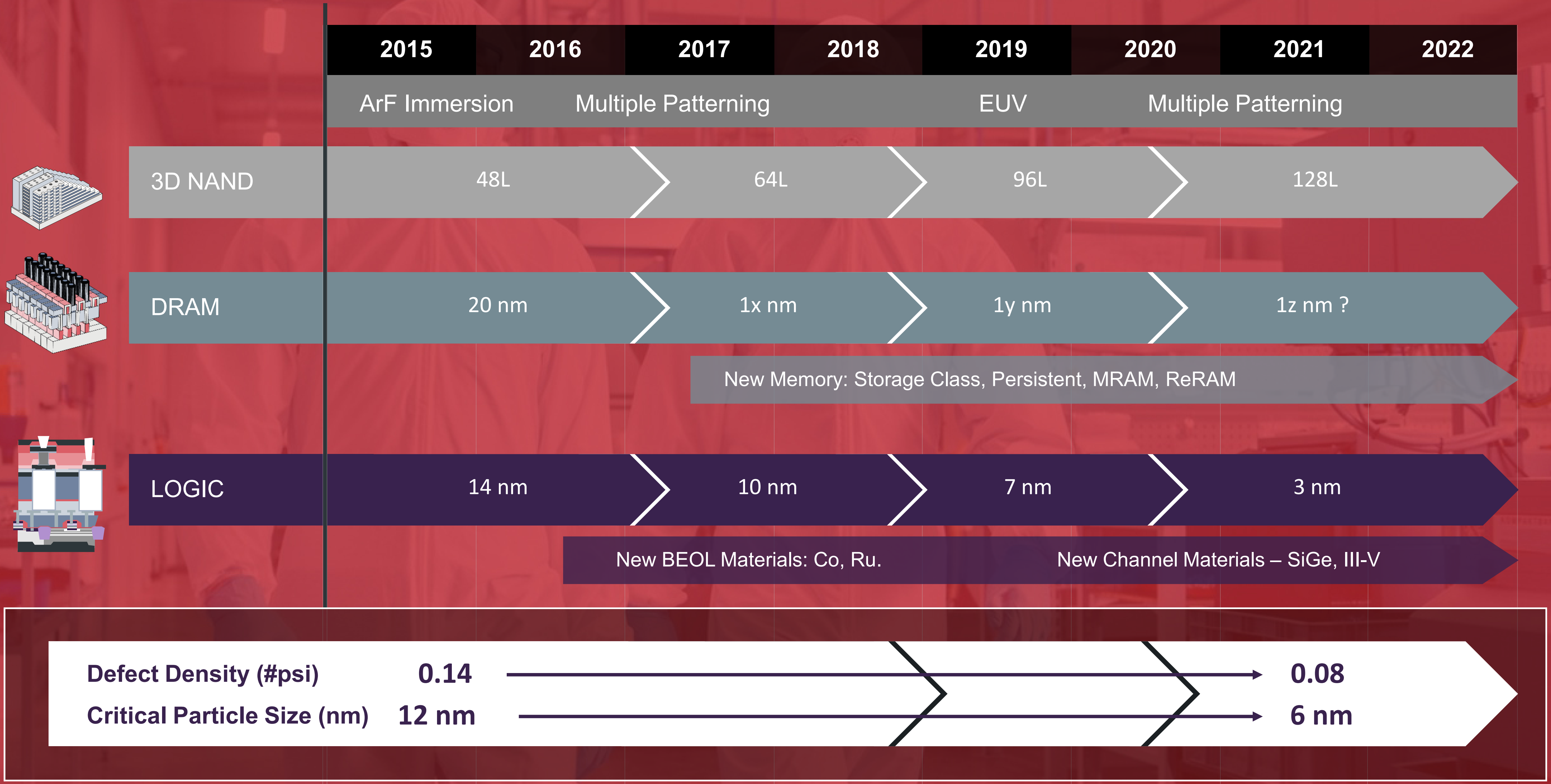
Introduction



As Moore's Law continues to manifest itself, it is clear that regardless of application, whether it's logic or memory, the feature size is shrinking and the number of stack layers is increasing, as gains in areal density become more and more expensive and semiconductor device manufacturers have gone vertical.

As a result, a need to create complex, three dimensional designs which pack more computing power and functionality into less space has emerged.
At the same time, the maximum tolerable defect size shrinks, as well as the maximum tolerable defect density in terms of number of particles per square inch.

Therefore new, innovative process technologies are required to accurately manufacture these complex 3D structures with minimal defects.



2015	2016	2017	2018	2019	2020	2021	2022

ArF Immersion Multiple Patterning EUV Multiple Patterning

3D NAND: 48L → 64L → 96L → 128L

DRAM: 20 nm → 1x nm → 1y nm → 1z nm ?

New Memory: Storage Class, Persistent, MRAM, ReRAM

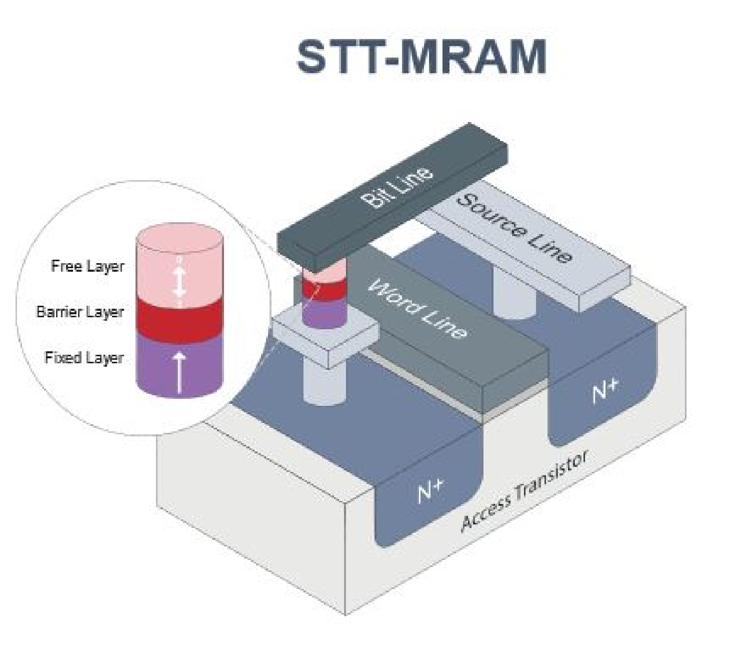
LOGIC: 14 nm → 10 nm → 7 nm → 3 nm

New BEOL Materials: Co, Ru. New Channel Materials – SiGe, III-V

Defect Density (#psi)	**0.14**	→	**0.08**
Critical Particle Size (nm)	**12 nm**	→	**6 nm**

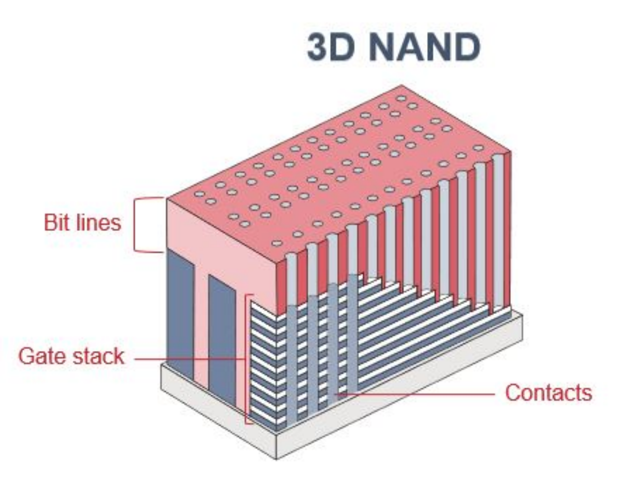
New challenges ahead

In these complex, multilayer structures, layers of alternating materials are deposited and then etched to form features. This etching process requires trenches to be dug through alternating layers of these differing materials, therefore the impact of any defects is very high and can result in a short of the device. These trenches then need to be filled and etched to form devices of the type you can see in the pictures.



+

Click here to see the pictures

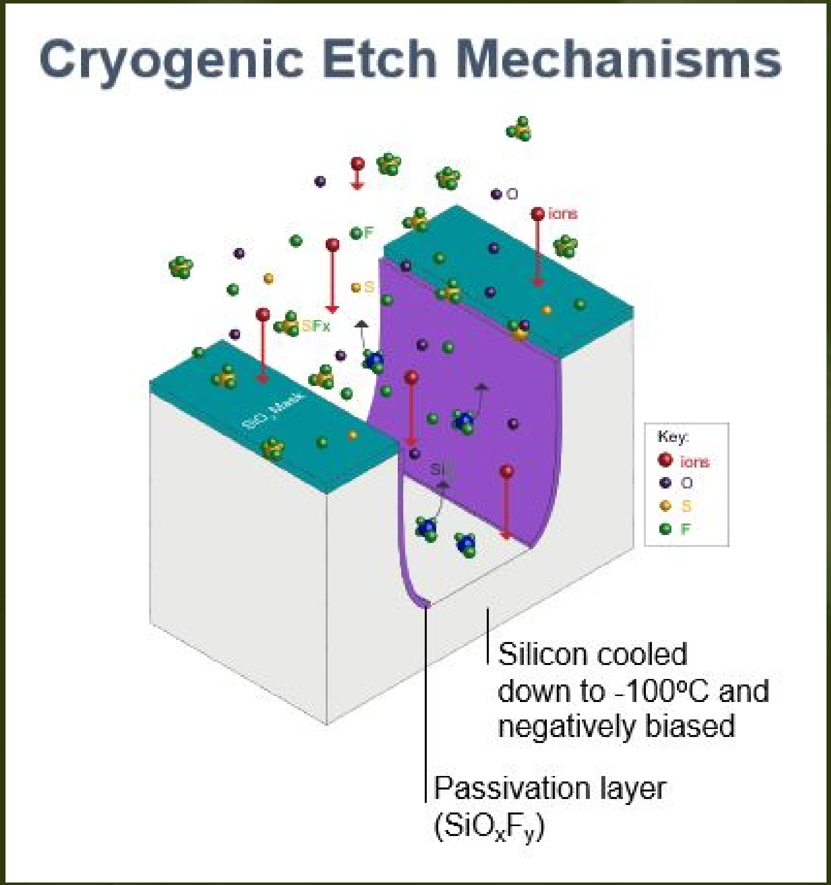


The problem is that the typical etch, for example at room temperature, is isotropic, meaning that it etches in all directions – but to dig these sorts of features and trenches you need directional etch. On top of that, due to the depth of these devices, the aspect ratio needs to be very high. So that's the high aspect ratio (HAR).

New challenges ahead

In order to achieve this, it is necessary to remove the material at the bottom, but not the sides. Removing material on the sides results in scalping of the trench wall, which is an undesirable effect.

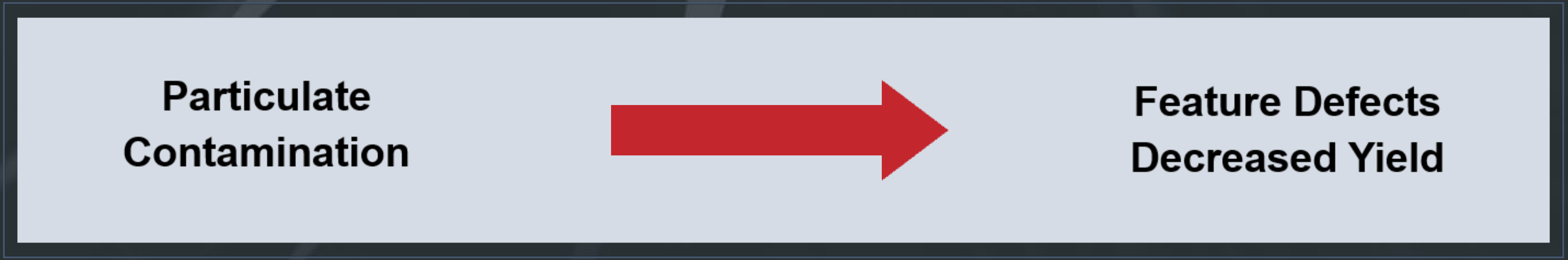
Cryogenic etch and cryogenic deposition are key enablers to form HAR devices at any kind of scale, and this requires that the substrates is cooled to very low temperatures, often below -100˚C.


Cryogenic Etch Mechanisms
O
ions
F
S
SiFx
SiO2 Mask
Si
Key:
ions
O
S
F
Silicon cooled down to -100ºC and negatively biased
Passivation layer (SiO_xF_y)

Contamination by particulates

In addition to the actual conditions to fabricate these devices, it is also necessary to minimise the potential for a defect incorporation, and therefore the need to minimise the possibility of contamination via particulates.

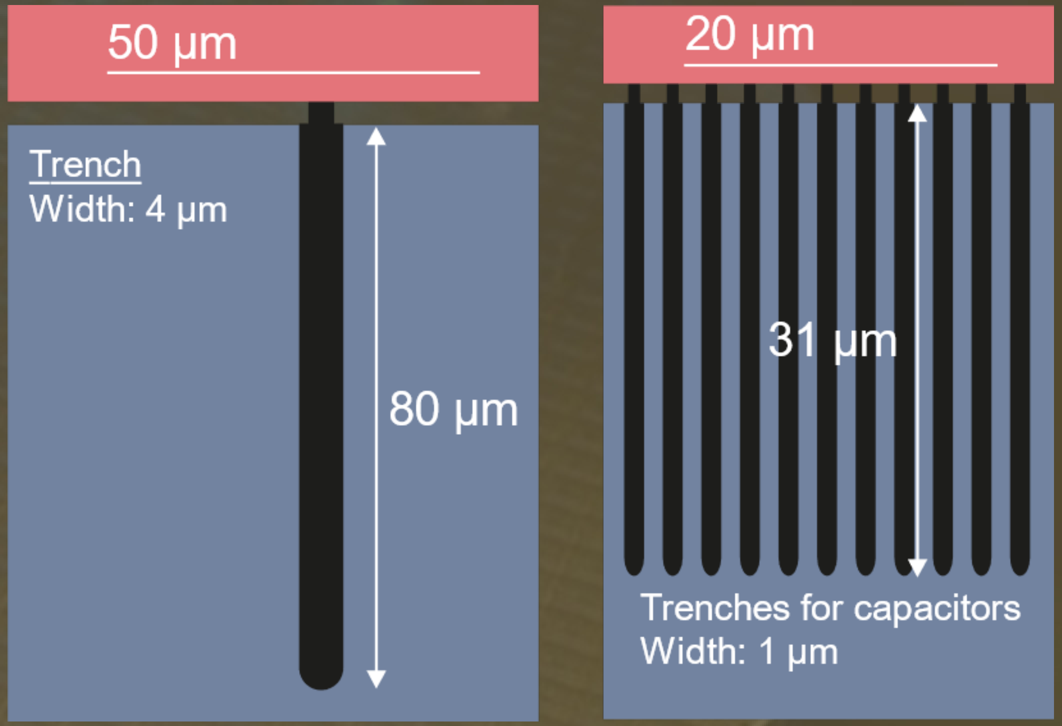
Particulate contamination is activated by the presence of H_2O, reacting with the fluorine in the plasma and resulting in AlF_3 particles – the number of particles generated is related to the partial pressure of water in the chamber. This is something that is very desirable to minimise, as we look at fabricating these very complex 3D structures.



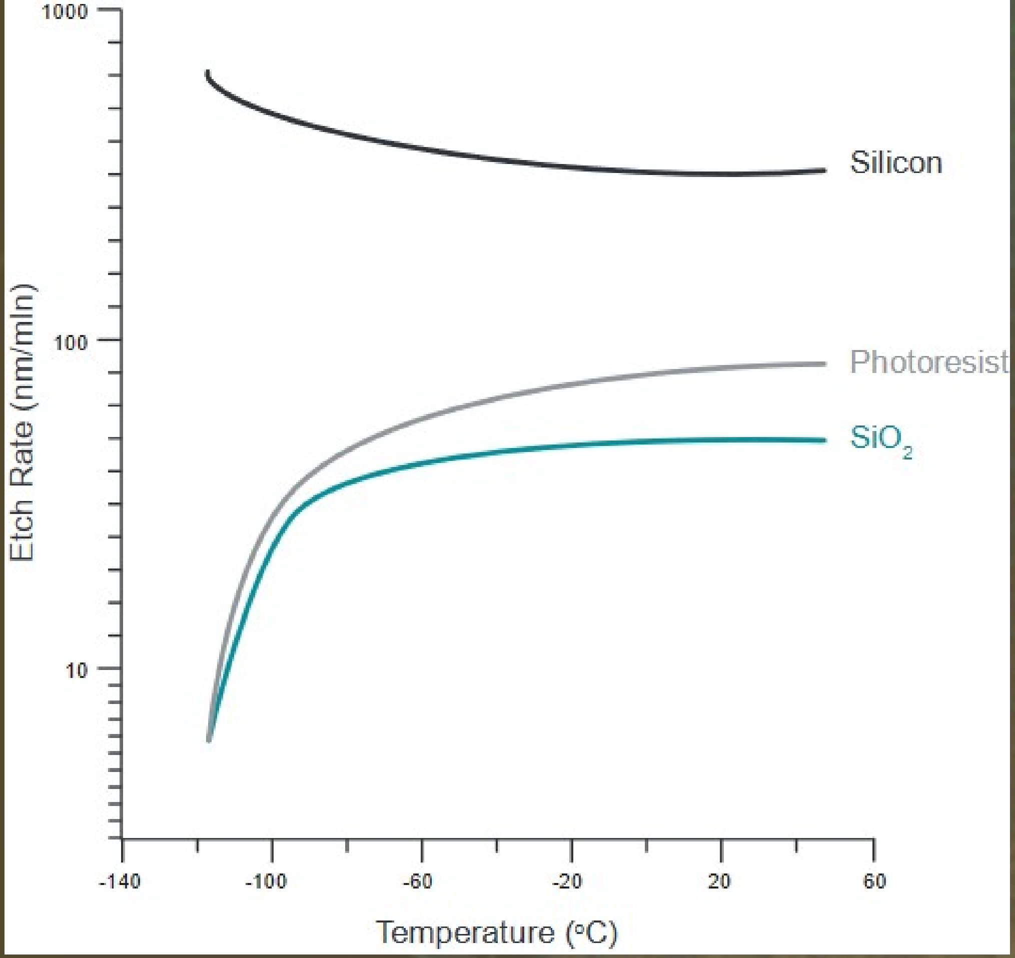
The advantages of cryogenic wafer cooling in etch processes

First of all, it is now possible to etch in a particular direction or anisotropically and create the very deep trenches required in this process.

You can get excellent sidewall passivation during the etch process, which minimises the scalloping effect because the SiO_xF_y passivation layer that is formed desorbs as you get to higher temperatures, meaning that it tends to stay on the device itself at lower temperatures reducing the number of process steps needed. Furthermore, at these low temperatures lateral diffusion of other species is also minimised. As a result, as the trenches are dug, the potential for damage in the lateral direction is minimised, together with the need to repair such damage.



The advantages of cryogenic wafer cooling in etch processes



Here you have a practical example: the etch rate for silicon, which is at the bottom of the trench, goes up as you reduce temperature, while the etch rate for the masking layer SiO_2 typically goes down as you go to lower temperatures, resulting in fewer etch defects at the top of the trench.

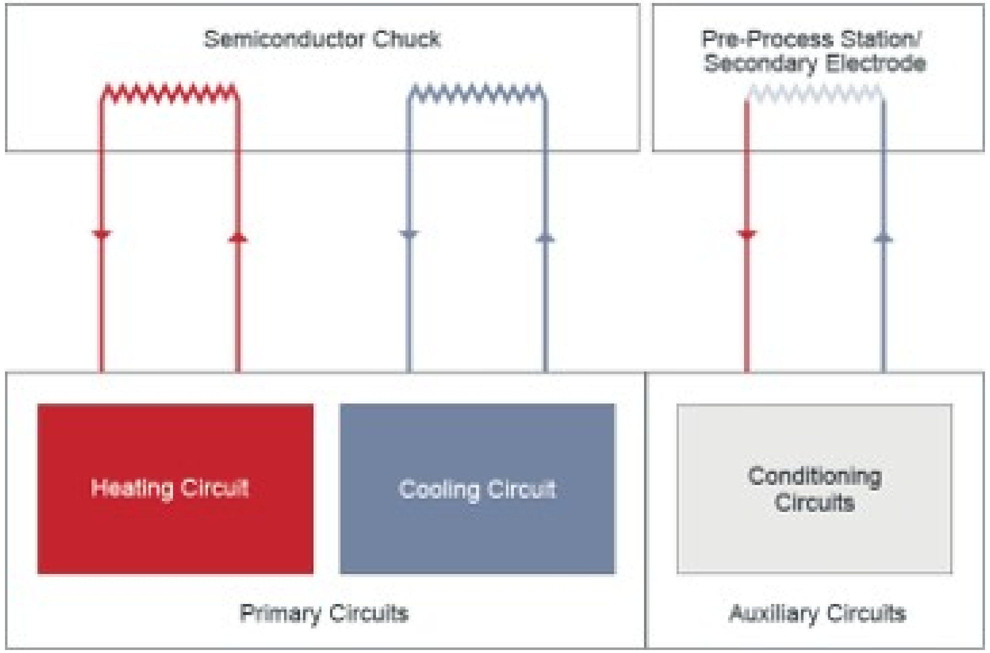
This combination drives selectivity towards deeper and narrower trenches, while preserving the integrity of the mask on the wafer surface.

How to achieve these cryogenic etch conditions using cooling systems

First of all, the cooling system may need an additional heating circuit in order to achieve not only the etch conditions, but also the annealing conditions for the wafer. This needs to be a closed-loop system to minimise operational costs.

Sometimes the chuck may be rotating rather than stationary, so the ability to handle this condition may also need to be engineered into the system.

The fact that the refrigeration fluid is actually going into the etch chamber where the plasma is present means that it needs to have high dielectric breakdown resistance, especially as in the etch process the energies of the plasma are fairly high.



We may also need to have multiple cooling and heating circuits to accommodate pre-process chambers prior to the main etch chamber.

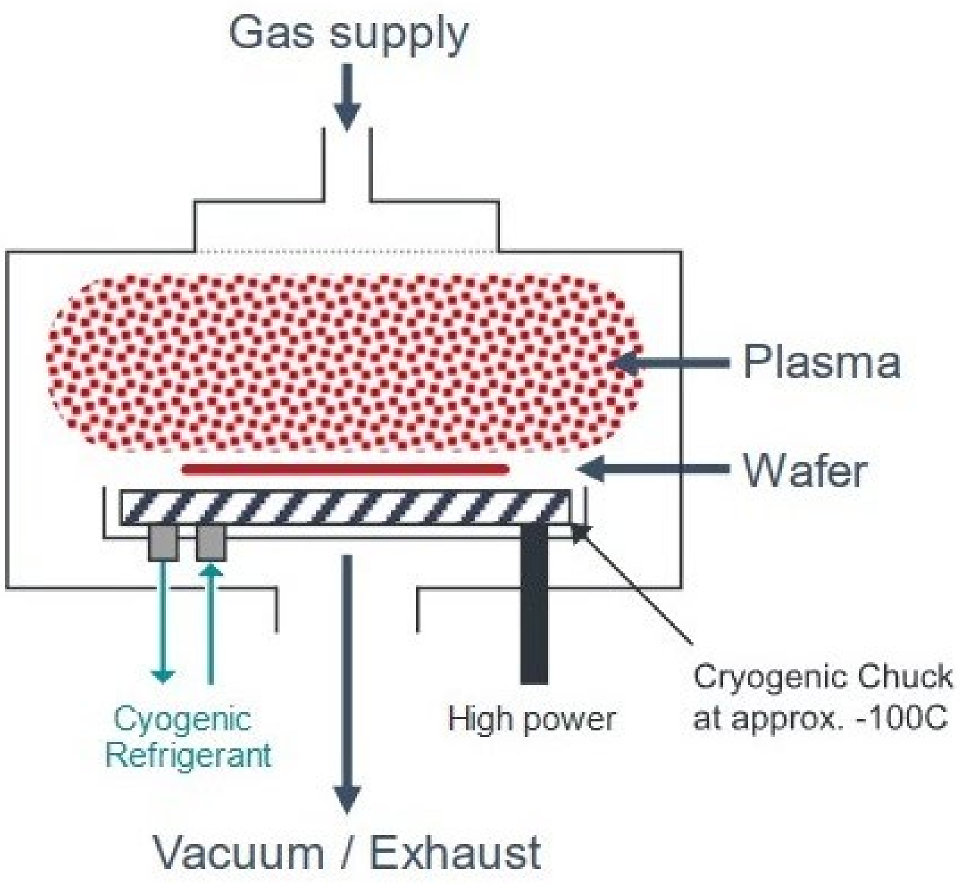
Finally, temperature control and temperature uniformity are very important. Temperature control is the ability to get to a cold or hot setpoint very quickly, within ±1^{o} C. Temperature uniformity is the ability to have uniformity across the wafer surface, so that the etch rates can be as homogeneous as possible across the area of the wafer – which also maximises the yield of these devices.

How to achieve these cryogenic etch conditions using cooling systems

As mentioned, in addition to target temperature, due to the plasma etching process the energy loads can also be fairly high:

- 8 to 15 kilowatts in the case of 3D NAND
- up to 2 kilowatts in the case of MRAM, albeit at lower temperatures

So whatever cooling equipment we use to attain the wafer temperature, it also needs to have enough heat capacity dynamically to remove the energy load from the chamber so as to not allow the wafer to deviate from its setpoint condition.

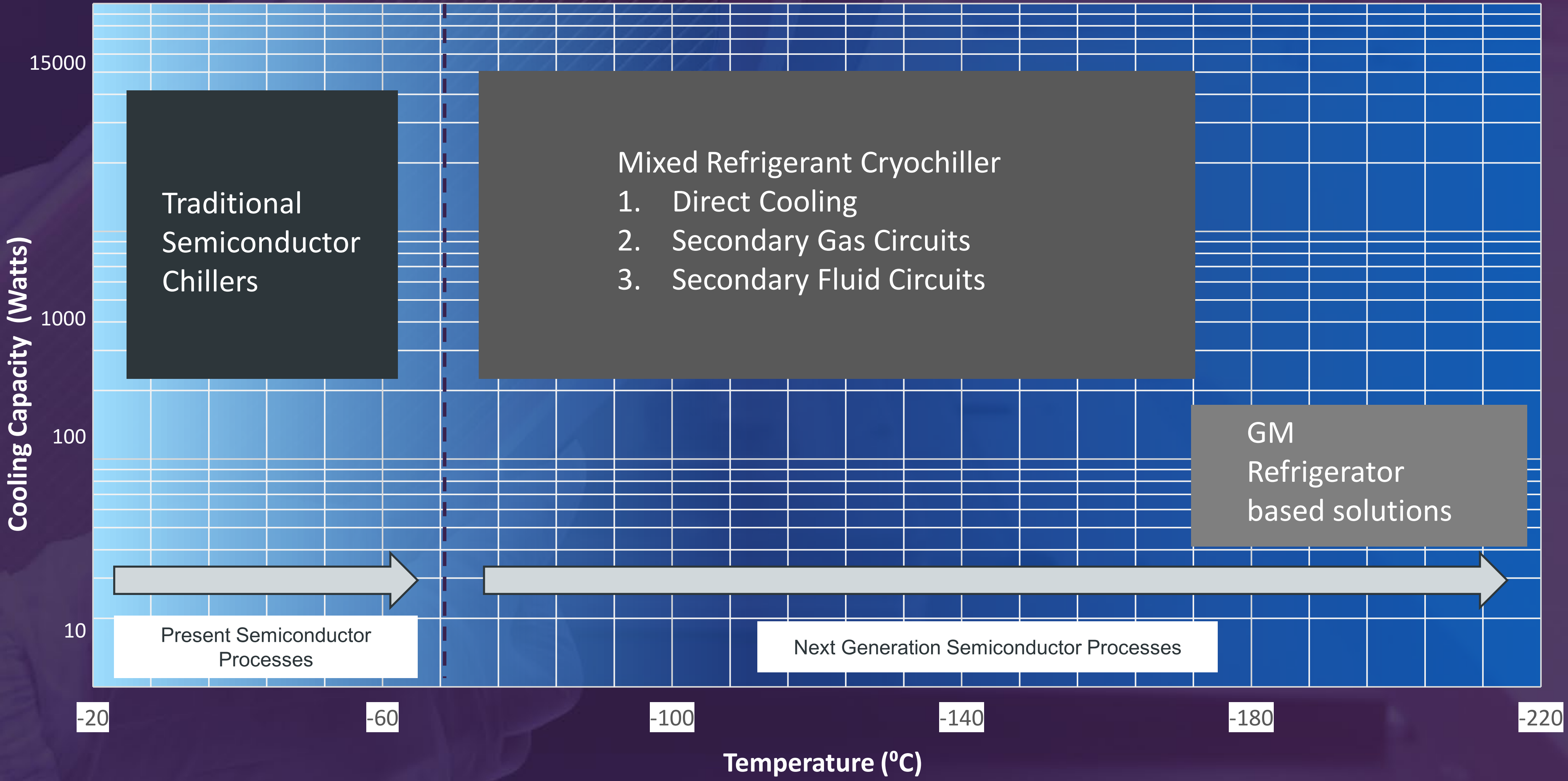


The challenges posed by the high demands on the cooling system used in cryogenic etch

Traditional chillers used in semiconductor manufacturing use glycol-based coolants and multi-stage compression to provide temperatures down to about -70˚C.

The next generation processes will require mixed refrigerant Joule-Thomson or MRJT cryochillers, which can deliver temperatures as low as -170˚C.

Beyond that, we will likely require a different technology, such as Gifford-McMahon refrigeration.



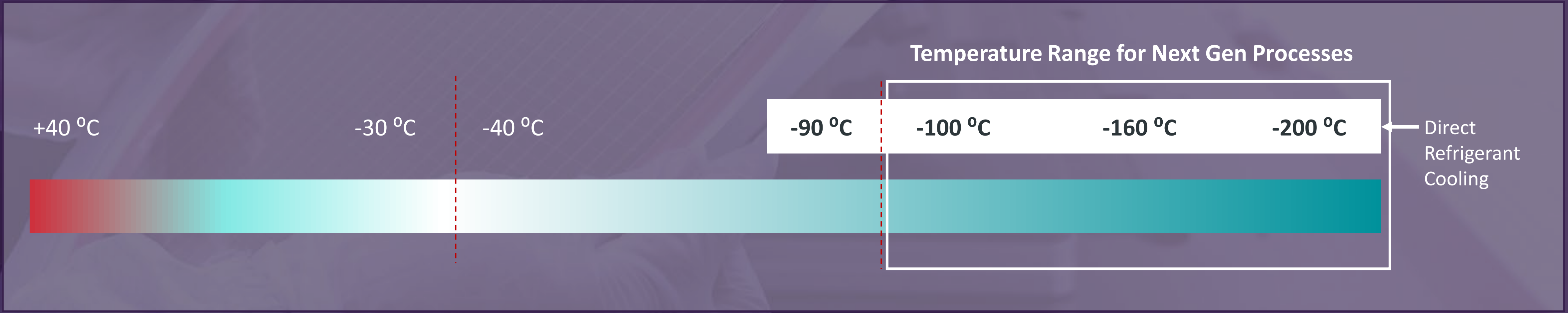
Cooling Capacity (Watts)
15000
1000
100
10
Traditional Semiconductor Chillers
Mixed Refrigerant Cryochiller
1. Direct Cooling
2. Secondary Gas Circuits
3. Secondary Fluid Circuits
GM Refrigerator based solutions
Present Semiconductor Processes
Next Generation Semiconductor Processes
-20
-60
-100
-140
-180
-220
Temperature (⁰C)

The challenges posed by the high demands on the cooling system used in cryogenic etch

Traditional glycol-based coolants are practical down to about -40˚C.

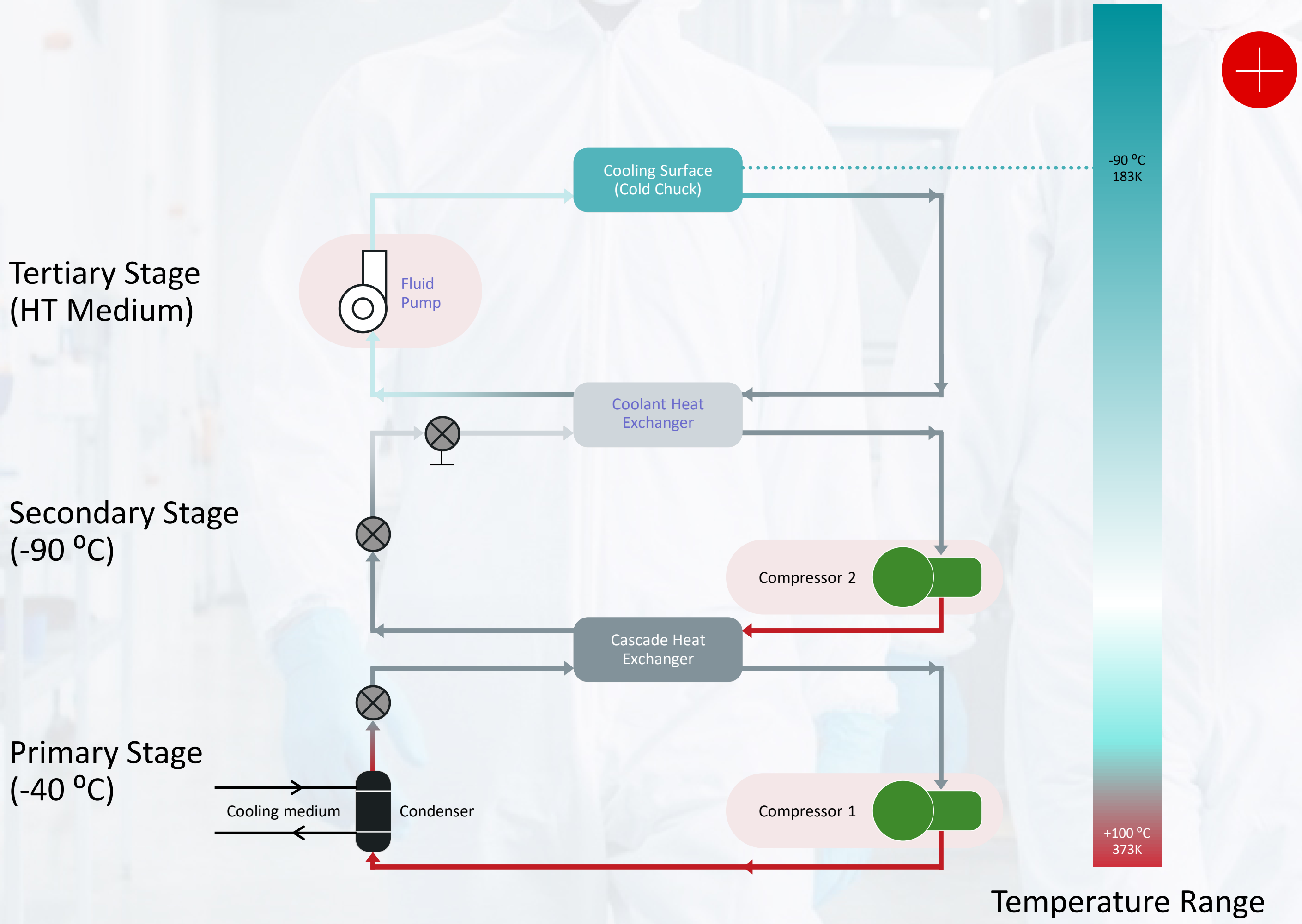
Advanced engineering coolants, which are often silicon-based and sometimes flammable, extend this range down to about -90˚C.

Below this temperature, it is no longer possible to use a dual fluid-to-fluid heat transfer mechanism – it is necessary to provide direct refrigerant cooling and to use refrigerants which can achieve much lower temperatures.


Temperature Range for Next Gen Processes
+40 °C
-30 °C
-40 °C
-90 °C
-100 °C
-160 °C
-200 °C
Direct Refrigerant Cooling

In addition, there are constraints around the dielectric breakdown, strength and material compatibility with the heat transfer channels and the chuck itself – and this is typically done by the OEM.

Direct refrigerant cooling – cascade chillers



A possible solution, albeit not the most efficient one, is represented by cascade chillers.

They use multiple stages and advanced thermal exchange media and can achieve temperatures as low as -90˚C, but often require multiple heat transfer mechanisms.

In addition, it might be necessary to implement multiple compressors and expansion cycles, leading to lower energy efficiency and greater spatial footprint.

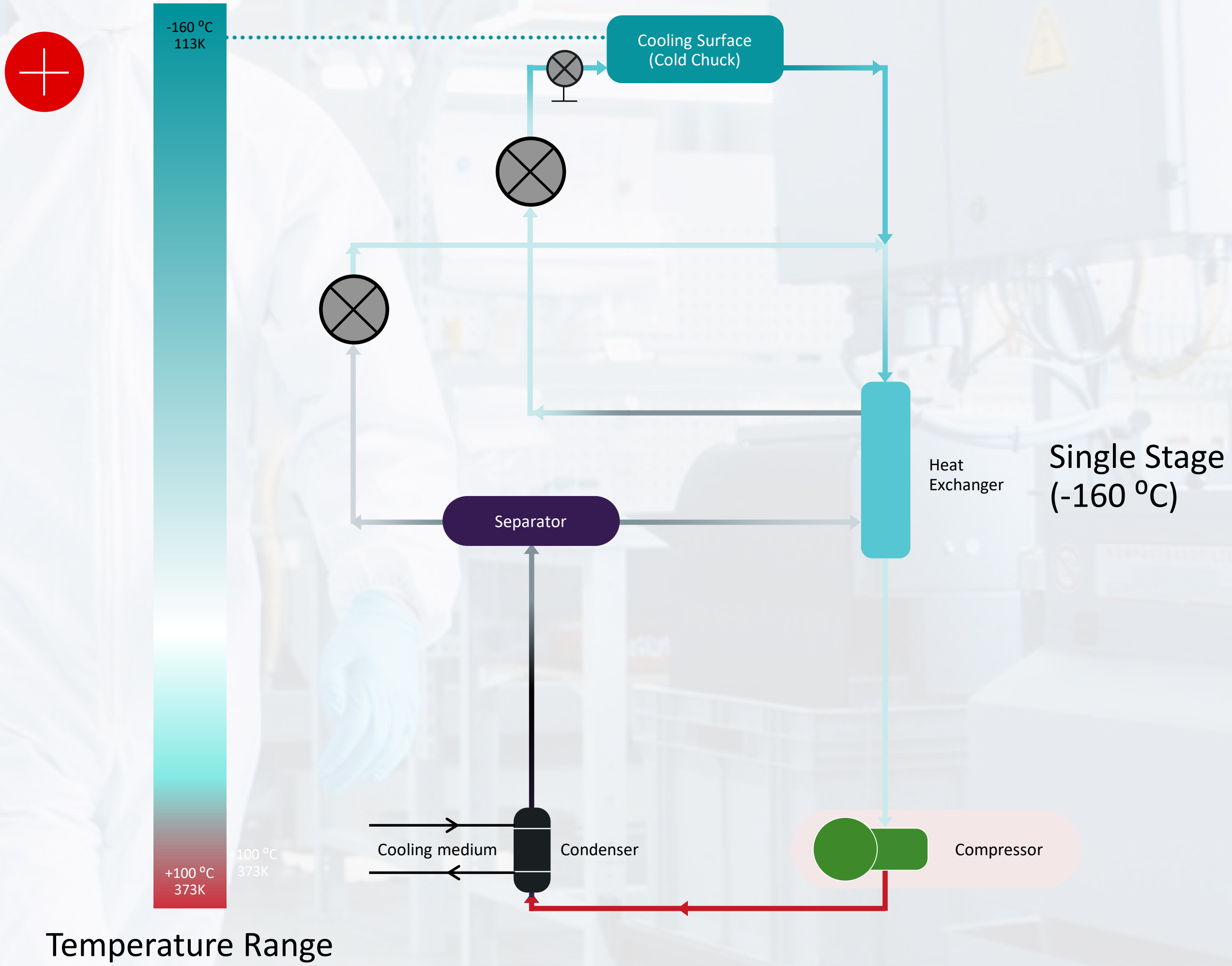
Direct refrigerant cooling – mix refrigerant cryochillers

A far more effective solution from a standpoint of both footprint and energy is the mixed refrigerant Joule-Thomson cryochiller, which uses a very similar process to what is found in a kitchen refrigerator or household air conditioner.

The working fluid is compressed, increasing its temperature and transferring heat to the external environment. The fluid is then allowed to expand through an orifice, cooling as it does so, in multiple stages, in what is called an auto cascade mechanism.

The mixed fluid is chosen specifically for thermodynamic properties such as enthalpy, entropy, and the phase diagram, and can be tailored to reach temperatures far lower than what is achievable with single refrigerants.

Flowing the refrigerant directly through the cooling surface of the chuck is far more effective than exchanging heat with a tertiary fluid, which then flows through the cooling surface of the chuck.

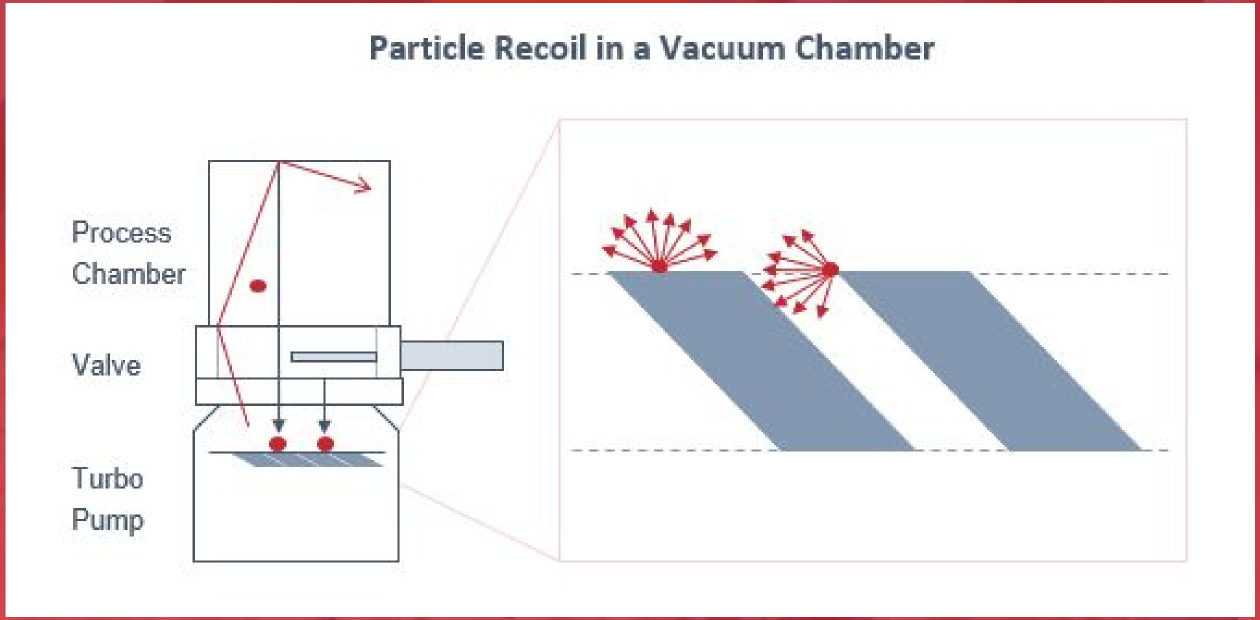


Defect reduction – how to have fewer particles in the chamber

Minimising the formation of particles is the main objective here, but what can be done once particles have already formed?

Our team at Edwards has developed models which show how the number of particles in the chamber is directly affected by the pumping mechanism – for example, up to 50% of the chamber particles can be recoiled back from the turbomolecular pump rotor back into the chamber, which is obviously not efficient.

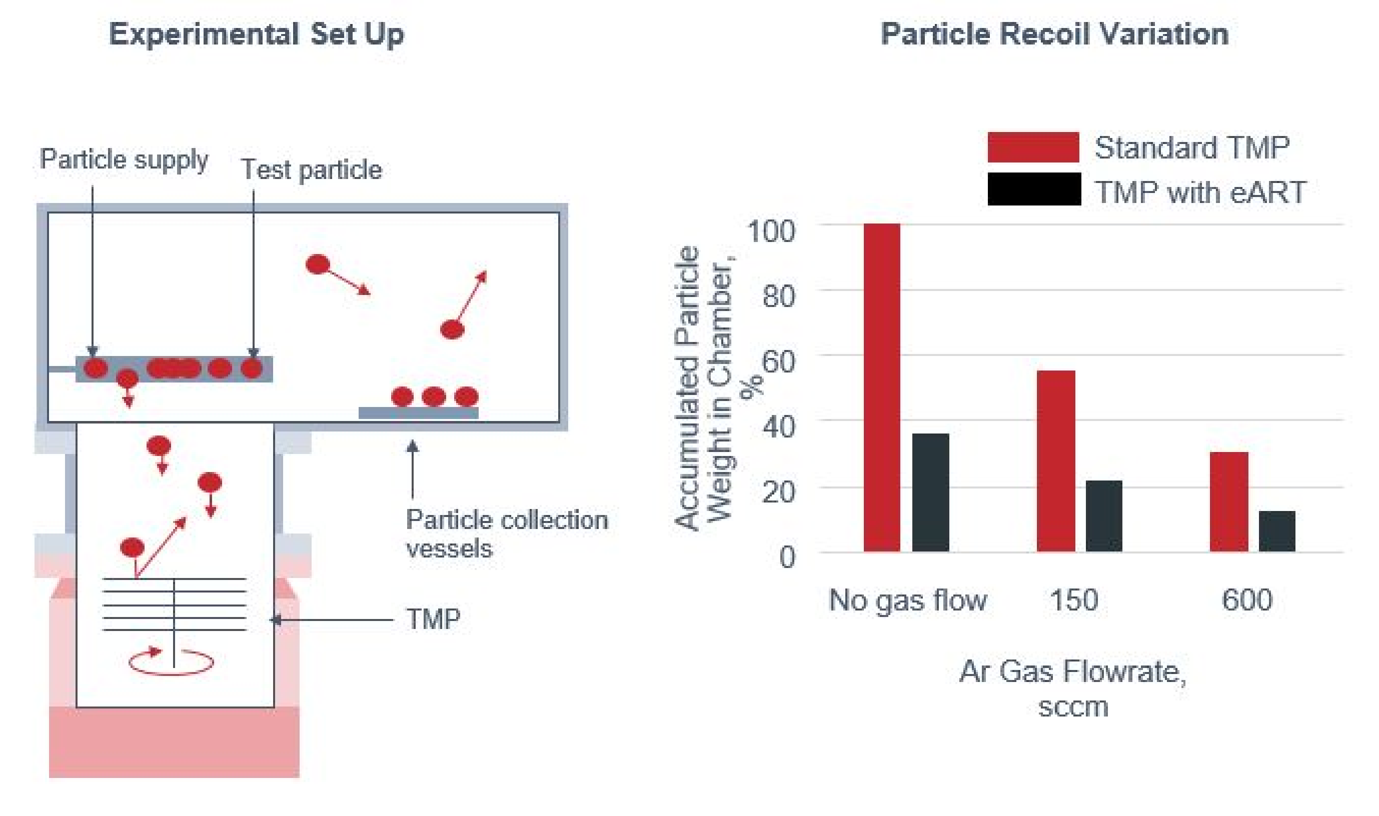
Luckily, there are a couple of ways to engineer this process more efficiently.



Defect reduction – how to have fewer particles in the chamber

We ran experiments where test particles from an artificial supply were introduced into the chamber environment, and the number of particles landing on a wafer-like collection device were counted in the presence or absence of certain design changes to the turbomolecular pump. Enhancements in materials such as low sticking coefficient and rotor geometry, which are patent protected, had significant influences.

We called the resulting combination the **Edwards Anti Recoil Technology (eART)**. A turbomolecular pump (TMP) implementing our eART technology shows significant reduction in accumulated particle weight in the chamber versus a standard TMP over a variety of argon gas flow rates. Once particles have formed, minimising the recoil back into the chamber is extremely important – and this is exactly what eART achieves.



Experimental Set Up

Particle Recoil Variation

Defect reduction – how to have fewer particles in the chamber

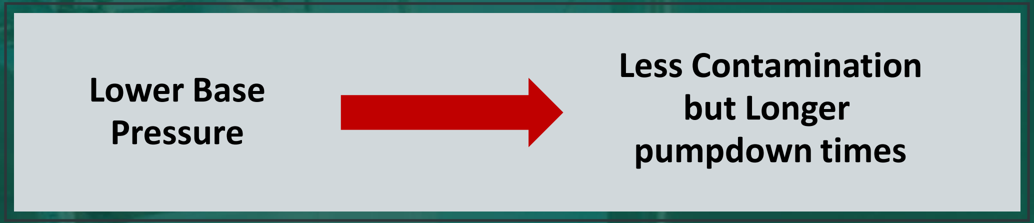
Watch the video of the process in action.

The particles enter the TMP, but instead of recoiling, thanks to our patented technology revolving around the geometry of the rotor blades, the number of particles passing through the rotor assembly and into the exhaust is higher than what is achieved with an ordinary TMP.

Defect reduction – how to minimise particle formation

One way to minimise particle formation is to reduce the water vapour in the chamber.

Firstly, this is needed to achieve the lower pressures in the mTorr range for the process. Water is notoriously difficult to pump compared to other gases because it tends to stick to surfaces rather than bounce off of them, on top of the fact that the porous ceramic surfaces in the chamber tend to absorb and retain high amounts of water vapour. In light of that, an effective way to remove water vapour from the chamber is very important just to attain the process pressures that you need.

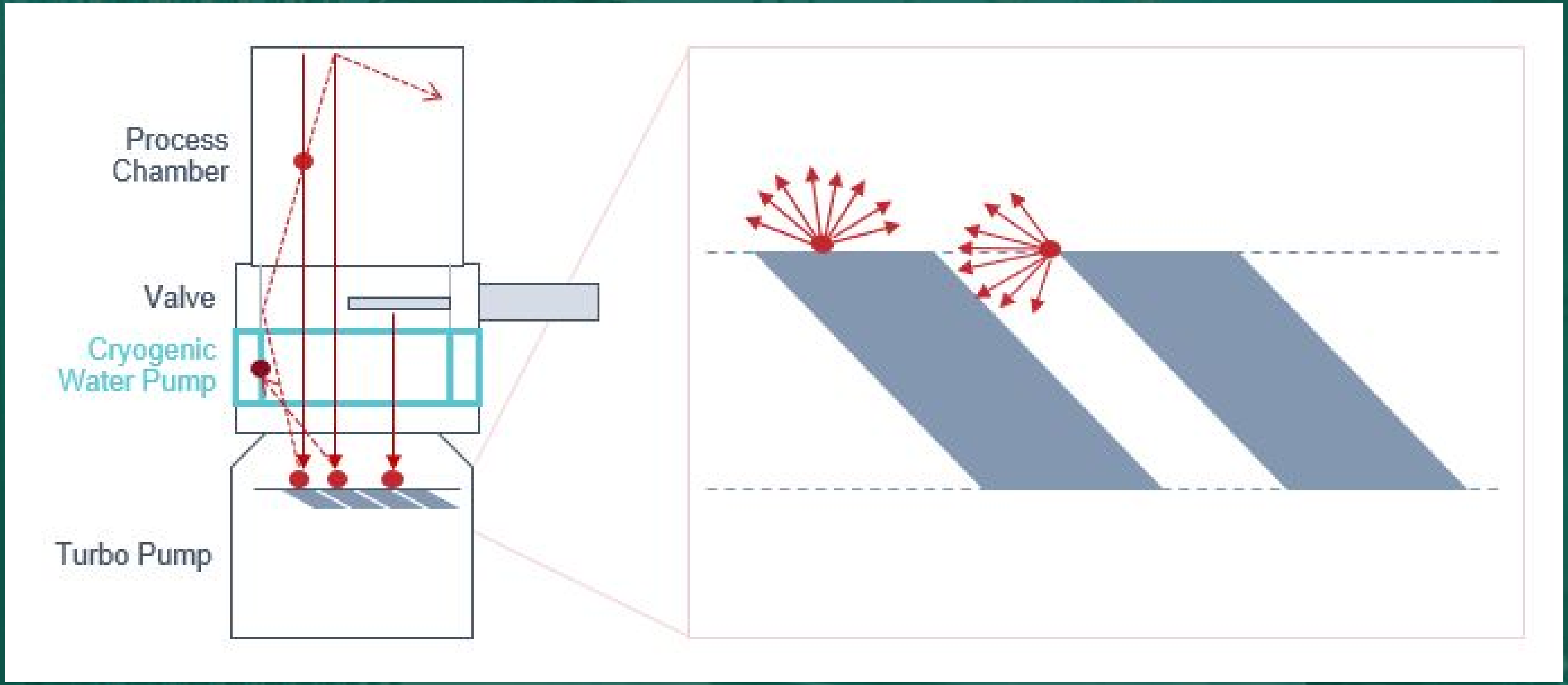

Lower Base Pressure → **Less Contamination but Longer pumpdown times**

In addition, the water reacts with fluorine and plasma to form HF. The HF then reacts with alumina to form AlF_3, regenerating the water – but AlF_3 is also a primary constituent of the residual particles in the chamber. Removing water therefore would not only allow us to get to the base chamber pressure needed, but also to remove one of the sources for the formation of AlF_3.

Defect reduction – how to minimise particle formation

We can do this effectively by adding a cryogenic water pump in series with the TMP. This is another trapping mechanism, similar to what is used in conventional cryopumps with cooling surfaces that will condense and retain water.

- ✓ Minimize H_2O partial pressure
- ✓ Lower base pressure
- ✓ Shorten cycle times
- ✓ Reduce particle counts
- ✓ Improve defectivity
- ✓ Increase yield



Conclusion

As semiconductor device designers and manufacturers continue to push more computing power and functionality into less space, they are moving into complex 3D structures with high aspect ratio features fabricated with advanced etch and deposition processes.

As shown, low temperatures are needed to achieve the directionality and selectivity in these etch processes. Mixed-refrigerant Joule-Thomson coolers can produce the necessary wafer temperatures and cooling power, especially in the case of high plasma energy processes with smaller footprint and lower power consumption than the conventional semiconductor chillers. But in addition to providing cooling to the wafer, we need to control particles, which is done in two ways.

First, we minimise the water vapour in the chamber, which helps base pressure as well, with a cryogenic water pump in series ✓

Then, we use a turbopump with advanced Anti-Recoil Technology to make sure that a high number of particles that have formed are captured and taken out through the exhaust, instead of them recoiling back into the chamber ✓

Together, these technologies will be critical enablers of the next generation processes and devices.

EDWARDS